Variable Screening Based on Combining Quantile Regression

by

Qian Shi

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

This thesis develops an efficient quantile-adaptive framework for linear and nonlinear variable screening with high-dimensional heterogeneous data.

Inspired by the success of various variable screening methods, especially in the quantile-adaptive framework, we develop a more efficient variable screening procedure. Both the classical linear regression model and the nonlinear regression model are investigated. In the thesis, the information over different quantile levels are combined, which can be implemented in two ways. The first one is the (weighted) average quantile estimator based on a (weighted) average of quantile regression estimators at single quantiles. The other one is the (weighted) composite quantile regression estimator based on a (weighted) quantile loss function.

Simulation studies are conducted to investigate the fine performance of the finite sample. A real data example is also analyzed.

Acknowledgements

First of all, I would like to express my sincerest gratitude to my supervisor Dr. Linglong Kong. During my study in the University of Alberta, he gave tremendous help and patient guidance to my study and research. He is such a responsible supervisor that he went through almost every small detail in my thesis. It would be much more difficult in finishing this thesis without his support.

I also wish to thank Dr. Narasimha Prasad, Dr. Ivan Mizera, and Dr. Ivor Cribben, for serving as members of my thesis examination committee.

It has been a truly enjoyable experience to work with the staff in the Department of Mathematical and Statistical Sciences, whose professional services are gratefully acknowledged.

Thanks to all my friends in Edmonton. During the two-year life at the University of Alberta, many friends have shown me their encouragement, given me their advice, and shared their experience with me. I will never forget the big time with them, and may our friendship be forever.

Finally, special thanks to my family, for their long-lasting understanding and support. I dedicate this thesis to them with love and gratitude.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Variable Selection

In recent years, high-dimensional data analysis has become increasingly frequent and important in a large variety of areas such as health sciences, economics, finance and machine learning. The analysis of high-dimensional data poses many challenges for statisticians and thus calls for new statistical methodologies as well as theories [6].

Variable selection plays an important role in high-dimensional statistical modeling. In practice, it is common to have a large number of candidate predictor variables available, and they are included in the initial stage of modeling for the consideration of removing potential modeling bias [5]. However, it is undesirable to keep irrelevant predictors in the final model, since it causes difficulty in interpreting the resultant model and may decrease its predictive ability.

There are many classical variable selection methods, for example, backward elimination, forward selection, stepwise selection, all subset selection and so on.

However, these algorithms often suffer from the high variability and may be trapped into a local optimal solution rather than the global optimal solution. Furthermore, if there are too many variables, classical variable selection may be very computationally expensive or even infeasible.

To deal with those problems, in the regularization framework, many different types of penalties have been successfully applied to achieve variable selection. Consider a sample $\left\{\left(\mathbf{x}_i, Y_i\right)^{\mathrm{T}}, i = 1, \ldots, n\right\}$ of size n from some unknown population, where $\mathbf{x}_i \in \mathbb{R}^p$. Taking, for example, the square loss function, we can select variables by solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + P_\lambda\left(\boldsymbol{\beta}\right),$$

where $P_\lambda\left(\boldsymbol{\beta}\right) = \sum_{j=1}^{p} p_\lambda\left(\beta_j\right)$ is a penalty function. A possible choice of the penalty function is the L_q penalty

$$L_q\left(\boldsymbol{\beta}\right) = \sum_{j=1}^{p} \left|\beta_j\right|^q, \quad q \geq 0.$$

For example, the L_2 penalty was introduced in ridge regression by Hoerl and Kennard [12]. The most popular and successful penalty is the L_1 penalty which was introduced for variable selection in the LASSO proposed by Tibshirani [27]. Its penalization term is given by

$$\lambda \sum_{j=1}^{p} \left|\beta_j\right|,$$

where $\lambda \geq 0$ is the regularization parameter. The LASSO continuously shrinks the coefficients toward 0 as λ increases with some coefficients shrunk to exactly

0 if λ is sufficiently large enough. Hence it can effectively select important variables and estimate regression parameters simultaneously. Under normal errors, the satisfactory finite sample performance of LASSO has been demonstrated numerically by Tibshirani [27], and its statistical properties have been studied by Knight and Fu [16], Fan and Li [5], and Tibshirani et al. [28]. However, the LASSO produces biased estimates for large coefficients, and thus it could be suboptimal in terms of estimation risk.

Fan and Li [5] argued that a good penalty should yield the following three properties in its estimator: unbiasedness, sparsity, and continuity. It is known that the L_q penalty with $q > 1$ does not satisfy the sparsity condition, whereas the L_1 penalty does not satisfy the unbiasedness condition, and the L_q penalty with $0 \leq q < 1$ does not satisfy the continuity condition. In other words, none of the L_q penalty family satisfies all three conditions simultaneously. For this reason, some penalties which satisfy those three conditions need to be proposed.

Zou [35] showed that there are scenarios in which the LASSO selection cannot be consistent. To cope with this problem, he proposed a new version of the LASSO, the adaptive LASSO, in which adaptive weights are used for penalizing different coefficients in the L_1 penalty. Specifically, he introduced the penalization term

$$\lambda \sum_{j=1}^{p} w_j \left|\beta_j\right|,$$

where $\lambda \geq 0$ is the regularization parameter and $\boldsymbol{w} = (w_1, \ldots, w_p)^{\mathrm{T}}$ is a known weight vector. A possible choice of weights can be derived from the estimator

of ordinary least squares regression $\tilde{\boldsymbol{\beta}}$:

$$\boldsymbol{w} = 1/\left|\tilde{\boldsymbol{\beta}}\right|.$$

The adaptive LASSO penalizes all coefficients consistently, and avoids possible biases associated with the LASSO. Therefore the adaptive LASSO enjoys the oracle properties, which were introduced by Fan and Li [5]: Let $\mathcal{A} = \left\{j : \beta_j^\star \neq 0\right\}$ be the true model and we further assume that $|\mathcal{A}| < p$. Thus the true model depends only on a subset of the predictors. The coefficient estimator produced by a fitting procedure δ is denoted by $\hat{\boldsymbol{\beta}}(\delta)$. Using the language of Fan and Li [5], we call δ an oracle procedure if $\hat{\boldsymbol{\beta}}(\delta)$ (asymptotically) has the following oracle properties:

- *Identifies the right subset model,* $\left\{j : \hat{\beta}_j \neq 0\right\} = \mathcal{A}$;
- *Has the optimal estimation rate,* $\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\delta)_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^\star\right) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}^\star)$*, where* $\boldsymbol{\Sigma}^\star$ *is the covariance matrix knowing the true subset model.*

Another popular penalty function which also shares the oracle properties was first introduced by Fan [3]. He proposed a nonconcave penalty function called the smoothly clipped absolute deviation (SCAD), which is defined by

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta|, & \text{if } |\beta| \leq \lambda \\ -\dfrac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{a(a-1)}, & \text{if } \lambda < |\beta| \leq a\lambda \\ \dfrac{(a+1)\lambda^2}{2}, & \text{if } |\beta| > a\lambda \end{cases},$$

where $a > 2, \lambda > 0$. The function is continuous and its first derivative can be

given by

$$p'_\lambda(|\beta|) = \lambda \left\{ I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right\} \qquad \text{for some } a > 2,$$

The SCAD corresponds to a quadratic spline function with knots at λ and $a\lambda$. For small coefficients, the SCAD is the same as LASSO, while it does not excessively penalize large values of $\boldsymbol{\beta}$. It also has the continuous solutions. In this way, SCAD achieves unbiasedness unlike LASSO. Fan and Li [5] suggested using $a = 3.7$ and showed that it has oracle properties in the penalized likelihood setting.

Recently a similar penalty called the minimax concave penalty (MCP) was introduced by Zhang [32]:

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| - \dfrac{\beta^2}{2\gamma}, & \text{if } |\beta| \leq \gamma\lambda \\ \dfrac{1}{2}\gamma\lambda^2, & \text{if } |\beta| > \gamma\lambda \end{cases}$$

for $\lambda > 0$ and $\gamma > 0$. The first derivative function of it is given by

$$p'_\lambda(\beta) = \begin{cases} \lambda - \dfrac{|\beta|}{\gamma}, & \text{if } |\beta| \leq \gamma\lambda \\ 0, & \text{if } |\beta| > \gamma\lambda. \end{cases}$$

The MCP as $\gamma \to \infty$ performs like the L_1 penalty. The MCP provides fast, continuous, nearly unbiased and accurate variable selection in high-dimensional linear regression.

1.2 Variable Screening

Although the variable selection methods mentioned above have been successfully applied to many high-dimentional analysis, the advent of modern technology for data collection pushes the dimensionality of data to a larger scale, that is we now encounter the situation where the dimensionality p is greater than the sample size n or even grows exponentially with n. The aforementioned variable selection methods may not work or perform well for these ultrahigh-dimentional data due to the simultaneous challenges of computational expediency, statistical accuracy and algorithm stability [8].

To address those challenges, a natural idea is to reduce the dimensionality p from a large or huge scale (say, $\log p = O(n^a)$ for some $a > 0$) to a relatively large scale d (e.g.,$O\left(n^b\right)$ for some $b > 0$) by a fast, reliable and efficient method, so that well-developed variable selection techniques can be applied afterwards. This provides a powerful tool for variable selection in ultrahigh-dimensional data analysis. It addresses the three issues, computational expediency, statistical accuracy and algorithm stability, as long as the variable screening procedure possesses the sure screening property introduced by Fan and Lv [7]. That is, *all truly important predictors can be selected with probability approaching one as the sample size goes to infinity.*

The two-scale method explicitly introduced by Fan and Lv [7] for ultrahigh-dimensional variable selection problems includes a crude large scale screening and a moderate scale selection, for example, the adaptive LASSO, the SCAD and the MCP. Usually, after the first step, the dimensionality will be reduced to less than the sample size n. They proposed sure independence screening (SIS) and iterated sure independence screening (ISIS), and showed that the

Pearson correlation ranking procedure possesses a sure screening property for linear regressions with Gaussian predictors and responses.

More specifically, suppose $\mathbf{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is the standardized response vector and $X_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}, j = 1, \ldots, p$, are the predictors. Let $\mathbf{X} = (X_1, \ldots, X_p)$ be the standardized predictor matrix. Suppose $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^{\mathrm{T}}$ is a p-vector that is obtained by componentwise regression, i.e.

$$\boldsymbol{\omega} = \mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Hence, $\boldsymbol{\omega}$ is essentially a vector of marginal correlations of predictors with the response variable.

For any given $\gamma \in (0, 1)$, Fan and Lv [7] sorted the p componentwise magnitudes of the vector $\boldsymbol{\omega}$ in a decreasing order and defined a submodel

$$\mathcal{M}_\gamma = \{1 \leq j \leq p : |\omega_j| \text{ is among the first } [\gamma n] \text{ largest of all}\},$$

where $[\gamma n]$ denotes the integer part of γn. This is a straightforward way to shrink the full model $\{1, \ldots, p\}$ down to a submodel $\mathcal{M}_\gamma$ with size $d = \mathcal{M}_\gamma < n$.

Independence screening means that each variable is used independently as a predictor to evaluate its usefulness for predicting the response. SIS is different from existing methods with penalization, as it does not use penalties to shrink the estimator. It ranks the importance of predictors according to their marginal correlation with the response variable and filters out the ones having weak marginal correlations with the response variable. Due to its independence screening property, the screening can be implemented very fast even

in the ultrahigh-dimentional case. Therefore, SIS has received a large amount of attention and has been further extended in various situations. For example, Fan, Samworth and Wu [8] extended ISIS to a general pseudo-likelihood framework, which includes generalized linear models as a special case. Fan and Song [9] proposed a more general version of the independence learning with ranking the maximum marginal likelihood estimators or the maximum marginal likelihood itself in generalized linear models. Fan, Feng and Song [4] considered nonparametric independence screening (NIS) in sparse ultrahigh-dimensional additive models. They suggested applying spline approximations to estimate the nonparametric components marginally, and ranking the importance of predictors based on the magnitude of the nonparametric components. This NIS procedure also possesses a sure screening property under some mild conditions.

Many other variable screening methods have emerged in the recent literature. Li, Peng and Zhang [20] proposed a robust rank correlation screening (RRCS) method based on a robust correlation Kendall τ. Li, Zhong and Zhu [22] developed a sure independence screening procedure based on the distance correlation (DC-SIS) under more general settings including linear models. Zhu et al. [34] introduced a model-free variable screening approach based on the relationship between each predictor and the indicator function $I(Y < y)$. There are many other methods appeared or to appear. The above lists are only the most relevant ones and not an attempt of a thorough review.

1.3 Variable Selection in Quantile Regression

Ordinary least squares regression estimates the mean response as a function of the predictors. As an alternative, least absolute deviation (LAD) regression estimates the conditional median function, which has been shown to be resistant to response outliers and more efficient when the errors have heavy tails. In the seminal paper of Koenker and Bassett [18], they generalized the idea of LAD regression and introduced quantile regression (QR) to estimate the conditional quantile function of the response. QR not only inherits the good properties of LAD regression but also provides much more information about the conditional distribution of the response variable.

The τth conditional function $Q_\tau(Y|X)$ is defined as

$$P\left(Y \leq Q_\tau(Y|X)\,|X = x\right) = \tau, \text{ for } 0 < \tau < 1.$$

By tilting the loss function, Koenker and Bassett [18] introduced the quantile check function which is defined by

$$\rho_\tau(u) = u\left(\tau - I(u < 0)\right).$$

They demonstrated that the τth conditional quantile function can be estimated by solving the following minimization problem

$$\min \sum_{i=1}^{n} \rho_\tau\left(Y_i - Q_\tau(Y|\mathbf{x}_i)\right). \tag{1.1}$$

Since its inception in Koenker and Bassett [18], QR has grown into an active research area in applied statistics.

To do variable selection in the quantile regression framework, we can also apply different penalties. For example, considering the penalized version of (1.1) in linear model, we solve the following optimization problem

$$\min \sum_{i=1}^{n} \rho_\tau \left(Y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) + P_\lambda \left(\boldsymbol{\beta}\right), \tag{1.2}$$

where $P_\lambda \left(\boldsymbol{\beta}\right) = \sum_{j=1}^{p} p_\lambda \left(\beta_j\right)$ is a penalty function. By choosing the appropriate penalty functions, variable selection can be achieved. Wang, Li and Jiang [29] proposed the LAD-LASSO, where τ equals to 0.5 and the LASSO penalty is applied in (1.2). Wu and Liu [31] focused on the variable selection based on penalized quantile regression, where the SCAD and the adaptive LASSO are employed in (1.2). Under some mild conditions, the oracle properties of the SCAD and the adaptive LASSO penalized quantile regression were demonstrated.

Partly inspired by the success of quantile regression, Zou and Yuan [36] considered composite quantile regression (CQR). For $\{\tau_k \in (0,1), k = 1, \ldots, K\}$, the estimator $\hat{\boldsymbol{\beta}}^{\mathrm{CQR}}$ is defined as

$$\left(\left\{\hat{b}^{\mathrm{CQR}}\left(\tau_k\right)\right\}_{k=1}^{K}, \hat{\boldsymbol{\beta}}^{\mathrm{CQR}}\right) = \arg\min_{b,\boldsymbol{\beta}} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k} \left(Y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - b\left(\tau_k\right)\right).$$

They also showed that the oracle model selection theory using the CQR oracle works beautifully even when the error variance is infinite. The CQR is an equally weighted sum of different quantile regressions at predetermined quantiles, which can be traced back to Koenker [17], who studied the estimator of

weighted averages of quantile regression objective functions in the form of

$$\left(\left\{\hat{b}^{\mathrm{WCQR}}\left(\tau_k\right)\right\}_{k=1}^{K}, \hat{\boldsymbol{\beta}}^{\mathrm{WCQR}}\right) = \arg\min_{b,\boldsymbol{\beta}} \sum_{k=1}^{K}\sum_{i=1}^{n} \varpi_k \rho_{\tau_k}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - b\left(\tau_k\right)\right).$$

where $\boldsymbol{\varpi} = \left(\varpi_1, \ldots, \varpi_K\right)^{\mathrm{T}}$ is a vector of weights and WCQR stands for weighted composite quantile regression. Intuitively, equal weights are not optimal in general. Koenker [17] as well as Zhao and Xiao [33] provided the optimal weight for WCQR to gain the most efficiency. Recently, Bradic, Fan and Wang [2] proposed a data-driven weighted linear combination of convex loss functions, together with weighted L_1 penalty method in the same spirit. As a specific example, they reintroduced the optimal composite quantile. Jiang, Jiang and Song [14] suggested using WCQR together with the adaptive LASSO and SCAD penalties in variable selection.

1.4 Quantile-Adaptive Variable Screening

In ultrahigh-dimensional data analysis, a new framework called quantile-adaptive sure independence screening (QaSIS) was proposed by He, Wang and Hong [11]. They advocated a quantile-adaptive approach which allows the set of active variables to be different when modeling various conditional quantiles. The screening method is based on the following observation:

$$Y \text{ and } X_j \text{ are independent } \Leftrightarrow Q_\tau\left(Y|X_j\right) - Q_\tau\left(Y\right) = 0, \quad \forall \tau \in (0,1),$$

where $Q_\tau\left(Y|X_j\right)$ is the τth conditional quantile of Y given the jth predictor X_j and $Q_\tau\left(Y\right)$ is the τth unconditional quantile of Y. In practice, the quantity

of the estimator of $Q_\tau(Y|X_j) - Q_\tau(Y)$ is expected to be close to zero if X_j is independent of Y. QaSIS is based on the magnitude of the estimator of $Q_\tau(Y|X_j) - Q_\tau(Y)$. In this quantile-adaptive model-free screening framework, $Q_\tau(Y|X_j)$ is estimated nonparametrically by B-spline approximations. In this aspect, this technique shares some similarity with NIS of Fan, Feng and Song [4] and Hall and Miller [10].

QaSIS provides a more complete picture of the conditional distribution of the response given all candidate predictors, and is more natural and effective in analyzing high-dimensional data especially those characterized by heteroscedasticity. Inherited from QR, QaSIS works well with heavy-tailed error distributions. Another distinctive feature of QaSIS is that it is model-free which avoids the specification of a particular model structure in a high-dimensional space.

1.5 Contributions of My Thesis

Motivated by the interesting work of He, Wang and Hong [11], we develop a more efficient variable screening procedure. Both the classical linear regression model and the nonlinear regression model are investigated. In the work of He, Wang and Hong [11], they considered only one quantile level. However, additional efficiency gains may be achieved by aggregating information over multiple quantiles. To combine information from multiple quantile regression, Zhao and Xiao [33] suggested two ways. The first one is to use an average of quantile regression estimators at each individual quantiles. The second one is to combine information over different quantiles via the criterion function. For example, CQR in Zou and Yuan [36] is along the second direction.

Following the two approaches in Zhao and Xiao [33], to combine information and improve the efficiency of the QaSIS method in He, Wang and Hong [11], we develop four estimators to screen variables. In more detail, the contributions of my thesis are summarized as follows:

- Propose four efficient estimators to screen variables, namely, the average quantile regression (AQR) estimator, the weighted average quantile regression (WAQR) estimator, the composite quantile regression (CQR) estimator and the weighted composite quantile regression (WCQR) estimator.

- Develop a screening procedure based on the above estimators in linear regression model and nonlinear regression model, where the B-spline approximations are employed in the latter case.

- Conduct simulation studies to investigate the finite sample performance of the screening procedure based on the four estimators in both linear and nonlinear models. Comparisons with other methods are also implemented.

- Propose the soft and hard threshold rules for the screening procedure and conduct simulation studies to study the two rules.

- Use the proposed methods to analyze a real data example, the large-B-cell lymphoma microarray data.

The rest of the thesis is organized as following. In Chapter 2, the new proposed screening methods based on the four estimators in both linear model and nonlinear model are introduced. Chapter 3 illustrates the finite sample per-

formance by both Monte Carlo simulations and a real data example analysis. Chapter 4 provides the summary and future work of my research.

Chapter 2

Variable Screening Based on Combining Quantile Regression

When the dimensionality p is high, say, $p > n$ or even grows exponentially with n, it is commonly assumed that only a small number of predictors among $X_1, \ldots, X_p$ actually contribute to the response Y, which leads to certain sparsity patterns in the unknown parameter $\boldsymbol{\beta}$. Our goal is to implement the first step of the two-scale method, namely the screening, to rapidly reduce the dimensionality p, usually greater than n, to a moderate scale via a computationally convenient procedure. We focus on the framework of quantile regression in this thesis. To deal with the heterogeneity in ultrahigh-dimensional data, He, Wang and Hong [11] advocated a quantile-adaptive sure independence screening (QaSIS) procedure. In particular, they assumed that at each quantile level only a sparse set of predictors is relevant for modeling Y, but allowed this set to be different at various quantiles, for instance, Example 4 in Chapter 3. Given a quantile level $\tau\,(0 < \tau < 1)$, $Q_\tau\,(Y|\mathbf{X})$ represents the τth conditional

quantile of Y given $\mathbf{X}$, that is

$$Q_\tau(Y|\mathbf{X}) = \inf\{y : P(Y \leq y|\mathbf{X}) \geq \tau\}.$$

The set of active variables is defined as

$$M_\tau = \{j : Q_\tau(Y|\mathbf{X}) \text{ functionally depends on } X_j\}.$$

Let $S_\tau = |M_\tau|$ be the cardinality of M_τ. Throughout this thesis, we assume $S_\tau, 0 < \tau < 1$, is smaller than the sample size n. Note that

$$Y \text{ and } X_j \text{ are independent } \Leftrightarrow Q_\tau(Y|X_j) - Q_\tau(Y) = 0, \quad \forall \tau \in (0,1), \quad (2.1)$$

where $Q_\tau(Y|X_j)$ is the τth conditional quantile of Y given the jth predictor X_j and $Q_\tau(Y)$ is the τth unconditional quantile of Y. Intuitively, one can see that, if X_j and Y are independent, adding the information of X_j does not change the quantile of Y, so that the conditional and unconditional quantile of Y, are the same. This is true for any $\tau \in (0,1)$.

2.1 Linear Model

Consider the problem of estimating a p-vector of parameters $\boldsymbol{\beta}$ in the following linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2.2)$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is a vector of responses, $\mathbf{X}$ is a $n \times p$ matrix of predictors with ith row $\mathbf{x}_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$ and jth column $X_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$,

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-vector of parameters and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ represents an n-vector of independent and identically distributed (i.i.d.) random errors, which is independent of $\mathbf{X}$.

To estimate the effect of X_j on the τth quantile of Y in practice, we consider the marginal quantile regression of Y on X_j, which can be expressed as follows:

$$\left(\hat{b}_j(\tau), \hat{\beta}_j(\tau)\right) = \arg\min_{b,\beta} \sum_{i=1}^{n} \rho_\tau (Y_i - x_{ij}\beta - b). \tag{2.3}$$

The minimization problem can be done easily using existing statistical software, for example, the *quantreg* package in R, *PROC QUANTREG* procedure in SAS and so on. Many algorithms, like the interior point method [25], simplex method [19] and MM algorithm [13] can also be easily implemented from scratch. Furthermore, we define

$$\widehat{D}_{nj}(\tau) = X_j \hat{\beta}_j(\tau) + \hat{b}_j(\tau) - \widehat{F}^{-1}_{Y,n}(\tau)$$

as an estimator of $D_{nj}(\tau) = Q_\tau(Y|X_j) - Q_\tau(Y)$, where $\widehat{F}^{-1}_{Y,n}(\tau)$ is the τth sample quantile function based on $Y_1, \ldots, Y_n$. According to (2.1), $\widehat{D}_{nj}$ measures the difference between the conditional and unconditional quantile of Y and is expected to be close to zero if X_j is independent of Y. One attractive advantage of $\widehat{D}_{nj}$ is that it does not need rescale either Y or X in practice, which may save some time and also make the results more interpretable.

The QaSIS is based on the magnitude of the estimated marginal components

$$\left\|\widehat{D}_{nj}\right\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}\hat{\beta}_j(\tau) + \hat{b}_j(\tau) - \widehat{F}^{-1}_{Y,n}(\tau)\right)^2.$$

To be more specific, we select the subset of variables

$$\widehat{M}_\tau = \left\{ 1 \leq j \leq p : \left\| \widehat{D}_{nj} \right\|_n^2 \geq \nu_n \right\},$$

where ν_n is a predefined threshold value. In sum, we propose to rank all the candidate predictors X_j, $j = 1, \ldots, p$ according to $\|\hat{D}_{nj}\|_n^2$ from the largest to smallest. We then select the top ones as the active predictors. Later we will propose several threshold rules to obtain the cutoff value that separates the active and inactive predictors, see Section 2.4.

It is possible that some slope coefficients are continuous and even constant in certain quantile intervals. The marginal quantile regression approach ignores such shared information across quantiles, only screens at individual quantile levels and thus may lose efficiency. A more efficient way is to combine information gained from different quantiles by considering the smoothness of the coefficient across quantiles. Suppose we want to consider K quantile levels, say $\{\tau_k \in (0,1), k = 1, \ldots, K\}$. A simple way is to take average of the estimated coefficients, which leads to the average quantile regression (AQR). Zhao and Xiao [33] argued that the simple average in general is not an efficient way of using distributional information from quantile regression. Moreover information at different quantiles are correlated, improperly using multiple quantile information may even reduce the efficiency, for example some effects may be balanced out. It is therefore important to combine quantile information appropriately to achieve more efficiency. Zhao and Xiao [33] studied the optimal combination of the estimated coefficients, which leads to the weighted average quantile regression (WAQR). When the coefficients across quantile levels are constant or approximately constant, the composite quantile regression (CQR)

which captures exactly this feature should be preferred. In addition, optimally combining the QR loss functions using weighted composite quantile regression (WCQR) will increase the efficiency furthermore.

2.1.1 Average Quantile Utility

Consider multiple quantile levels $\{\tau_k \in (0,1), k = 1, \ldots, K\}$. For each τ_k, we estimate β via the marginal quantile regression and define the average quantile regression (AQR) estimator as

$$\hat{\beta}_j^{\mathrm{AQR}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\beta}_j(\tau_k), \qquad \hat{b}_j^{\mathrm{AQR}} = \frac{1}{K}\sum_{k=1}^{K} \hat{b}_j(\tau_k),$$

where $\hat{\beta}_j(\tau_k)$ and $\hat{b}_j(\tau_k)$ are defined in (2.3) at the quantile level τ_k.

Since the equivalent relationship (2.1) holds for any given $\tau \in (0,1)$, $D_{nj}^{\mathrm{AQR}} = K^{-1}\sum_{k=1}^{K}\left[Q_{\tau_k}(Y|X_j) - Q_{\tau_k}(Y)\right]$ is expected to be close to zero if X_j is independent of Y, as it measures the information that X_j brings in to estimate the quantiles of Y. If X_j contributes to the quantiles of Y only at several quantile levels, it can also be captured by D_{nj}^{AQR}.

We define

$$\begin{aligned}\widehat{D}_{nj}^{\mathrm{AQR}} &= X_j\hat{\beta}_j^{\mathrm{AQR}} + \hat{b}_j^{\mathrm{AQR}} - \frac{1}{K}\sum_{k=1}^{K}\widehat{F}_{Y,n}^{-1}(\tau_k) \\ &= \frac{1}{K}\sum_{k=1}^{K}\left(X_j\hat{\beta}_j(\tau_k) + \hat{b}_j(\tau_k) - \widehat{F}_{Y,n}^{-1}(\tau_k)\right)\end{aligned}$$

as an estimator of D_{nj}^{AQR}. $\widehat{D}_{nj}^{\mathrm{AQR}}$ is expected to be close to zero if X_j is independent of Y. It can capture the information that X_j brings in to estimate the quantiles of Y, even if X_j contributes to the quantiles of Y only at several

quantile levels.

The independence screening is also based on the magnitude of the estimated marginal components, which is given by

$$\begin{aligned}\left\|\widehat{D}_{nj}^{\mathrm{AQR}}\right\|_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[x_{ij}\hat{\beta}_j^{\mathrm{AQR}} + \hat{b}_j^{\mathrm{AQR}} - \frac{1}{K}\sum_{k=1}^{K}\widehat{F}_{Y,n}^{-1}(\tau_k)\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{K}\sum_{k=1}^{K}\left(x_{ij}\hat{\beta}_j(\tau_k) + \hat{b}_j(\tau_k) - \widehat{F}_{Y,n}^{-1}(\tau_k)\right)\right]^2.\end{aligned}$$

Finally we select the following subset of variables by the threshold of $\|\widehat{D}_{nj}^{\mathrm{AQR}}\|_n^2$:

$$\widehat{M}^{\mathrm{AQR}} = \left\{1 \le j \le p : \left\|\widehat{D}_{nj}^{\mathrm{AQR}}\right\|_n^2 \ge \nu_n\right\},$$

where ν_n is a predefined threshold value.

In practice, we only need to choose a relatively small K, say $K = 9$ [2] or 19 [36]. Throughout this thesis, we fixed $K = 9$. Since each variable is used independently as a predictor to decide its usefulness for predicting the response, the screening can be done very fast using existing software even in the ultrahigh-dimensional case.

2.1.2 Weighted Average Quantile Utility

Regardless of the actual amounts of contributions from quantiles and the possible correlations between them, the AQR method puts the equal weight on different quantile levels. In this way, it is likely that the effects from different quantiles can be balanced out. Therefore, it may be more effective if appropriate weights are applied by considering the possible variation in contributions and correlations between multiple quantiles, which leads to the weighted av-

erage quantile regression (WAQR) estimator:

$$\hat{\beta}_j^{\mathrm{WAQR}} = \sum_{k=1}^{K} \omega_k \hat{\beta}\left(\tau_k\right), \qquad \hat{b}_j^{\mathrm{WAQR}} = \sum_{k=1}^{K} \omega_k \hat{b}\left(\tau_k\right), \tag{2.4}$$

where $\hat{\beta}_j\left(\tau_k\right)$ and $\hat{b}_j\left(\tau_k\right)$ are defined in (2.3) at the quantile level τ_k, $\boldsymbol{\omega} = \left(\omega_1, \ldots, \omega_K\right)^{\mathrm{T}}$ is a vector of weights. The weight ω_k controls the contribution of the τ_kth QR and possible correlations. In AQR, all the weights are $1/K$. For WAQR, we can choose an optimal weighting scheme to obtain the most efficiency.

Koenker [17] provided the optimal weights for WAQR by minimizing the variance of the estimator. Suppose that the error ε has the distribution function F with density f, such that $0 < f\left(F^{-1}\left(\tau\right)\right) < \infty$ for $\tau \in \{\tau_1, \ldots, \tau_K\}$. Throughout this thesis, we use the vector $\mathbf{v}$ and matrices B and V defined as follows:

$$\begin{aligned}
&\mathbf{v} = \left(v_1, \ldots, v_K\right)^{\mathrm{T}}, \quad \text{where } v_k = f\left(F^{-1}\left(\tau_k\right)\right), \\
&V = \mathrm{diag}(\mathbf{v}), \\
&B = \left[\min\left(\tau_i, \tau_j\right) - \tau_i \tau_j\right]_{1 \le i,j \le k}.
\end{aligned}$$

In Theorem 5.2 (page 169) of Koenker [17], under some conditions, as the sample size $n \to \infty$,

$$\sqrt{n}\left(\hat{\beta}_j^{\mathrm{WAQR}} - \beta\right) \to N\left(0, J\left(\boldsymbol{\omega}\right) \boldsymbol{\Sigma}_X^{-1}\right),$$

where $\mathbf{\Sigma}_{X_j}$ is the covariance of X_j, which is fixed and

$$J(\boldsymbol{\omega}) = \boldsymbol{\omega}^{\mathrm{T}} V^{-1} B V^{-1} \boldsymbol{\omega} \geq \left(\mathbf{v}^{\mathrm{T}} B \mathbf{v}\right)^{-1}. \tag{2.5}$$

The covariance matrix of $\hat{\beta}_j^{\mathrm{WAQR}}$ depends on the weights through $J(\boldsymbol{\omega})$. Thus a natural way to select optimal weights in the WAQR is to minimize $J(\boldsymbol{\omega})$ in (2.5) to get the smallest variance of the estimator. The optimal weight $\boldsymbol{\omega}_{\mathrm{opt}}$ can be obtained by some simple algebra calculation as follows:

$$\boldsymbol{\omega}_{\mathrm{opt}} = \left(\mathbf{v}^{\mathrm{T}} B \mathbf{v}\right)^{-1} V B^{-1} \mathbf{v}. \tag{2.6}$$

If X_j and Y are independent, then $D_{nj}^{\mathrm{WAQR}} = \sum_{k=1}^{K} \omega_k \left[Q_{\tau_k}(Y|X_j) - Q_{\tau_k}(Y)\right]$ is supposed to be close to zero, as it measures the information that X_j brings in to estimate the quantiles of Y for all the selected quantile levels. We define the sample version of D_{nj}^{WAQR}

$$\begin{aligned}
\widehat{D}_{nj}^{\mathrm{WAQR}} &= X_j \hat{\beta}_j^{\mathrm{WAQR}} + \hat{b}_j^{\mathrm{WAQR}} - \sum_{k=1}^{K} \omega_{\mathrm{opt},k} \widehat{F}_{Y,n}^{-1}(\tau_k) \\
&= \sum_{k=1}^{K} \omega_{\mathrm{opt},k} \left(X_j \hat{\beta}_j(\tau_k) + \hat{b}_j(\tau_k) - \widehat{F}_{Y,n}^{-1}(\tau_k)\right).
\end{aligned}$$

As mentioned before, if X_j and Y are independent, then $\widehat{D}_{nj}^{\mathrm{WAQR}}$ is supposed to be close to zero, as it can capture all the information that X_j brings in to estimate the quantiles of Y for all the quantile levels. If X_j is an important predictor variable, then the difference between the conditional and unconditional quantiles of Y is expected to be away from zero.

We rank and screen the variables based on the magnitude of the estimated

marginal components

$$\begin{aligned}\left\|\widehat{D}_{nj}^{\mathrm{WAQR}}\right\|_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[x_{ij}\hat{\beta}_j^{\mathrm{WAQR}} + \hat{b}_j^{\mathrm{WAQR}} - \sum_{k=1}^{K}\omega_{\mathrm{opt},k}\widehat{F}_{Y,n}^{-1}\left(\tau_k\right)\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}\omega_{\mathrm{opt},k}\left(x_{ij}\hat{\beta}_j\left(\tau_k\right) + \hat{b}_j\left(\tau_k\right) - \widehat{F}_{Y,n}^{-1}\left(\tau_k\right)\right)\right]^2,\end{aligned} \tag{2.7}$$

and we focus on the following subset of variables

$$\widehat{M}^{\mathrm{WAQR}} = \left\{1 \le j \le p : \left\|\widehat{D}_{nj}^{\mathrm{WAQR}}\right\|_n^2 \ge \nu_n\right\},$$

where ν_n is a predefined threshold value. This screening procedure is fast since each variable is used as a predictor to decide the usefulness for predicting Y.

2.1.3 Composite Quantile Utility

To combine the information from multiple quantile levels, besides the AQR (WAQR) method, another frequently applied approach is to use composite quantile regression (CQR). It uses the information that the coefficients are constant explicitly in the model. It combines information over different quantiles via a different criterion function. We estimate the model over $\{\tau_k \in (0,1), k = 1, \ldots, K\}$ jointly based on the following modified quantile loss function:

$$\left(\left\{\hat{b}_j^{\mathrm{CQR}}\left(\tau_k\right)\right\}_{k=1}^{K}, \hat{\beta}_j^{\mathrm{CQR}}\right) = \arg\min_{b,\beta}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}\left(Y_i - x_{ij}\beta - b\left(\tau_k\right)\right).$$

Typically, we use equally spaced quantiles: $\tau_k = \frac{k}{K+1}$ for $k = 1, \ldots, K$. Note that the regression coefficients remain the same across different quantile re-

gressions. This minimization problem can be solved by the MM algorithm, but unfortunately it cannot be done directly by any software. In practice, we implement the algorithm in R.

As the equivalent relationship (2.1) holds for any given $\tau \in (0,1)$, if X_j and Y are independent, $D_{nj}^{\mathrm{CQR}} = \sum_{k=1}^{K} \left[Q_{\tau_k}(Y|X_j) - Q_{\tau_k}(Y) \right]$ is expected to be close to zero. Since the estimated coefficients over different quantiles are the same, D_{nj}^{CQR} measures the entire change between the unconditional and conditional quantiles of Y among all the quantiles. If X_j is the important predictor variable, D_{nj}^{CQR} should not be close to zero.

We define

$$\widehat{D}_{nj}^{\mathrm{CQR}} = \sum_{k=1}^{K} \left(X_j \hat{\beta}_j^{\mathrm{CQR}} + \hat{b}_j^{\mathrm{CQR}}(\tau_k) - \widehat{F}_{Y,n}^{-1}(\tau_k) \right)$$

as an estimator of D_{nj}^{CQR}. As the equivalent relationship (2.1) holds for any given $\tau \in (0,1)$, then $\widehat{D}_{nj}^{\mathrm{CQR}}$ is supposed to be close to zero if X_j and Y are independent.

Then we rank the predictors by the magnitude of the estimated marginal components

$$\left\| \widehat{D}_{nj}^{\mathrm{CQR}} \right\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \left(x_{ij} \hat{\beta}_j^{\mathrm{CQR}} + \hat{b}_j^{\mathrm{CQR}}(\tau_k) - \widehat{F}_{Y,n}^{-1}(\tau_k) \right) \right]^2 .$$

Finally, the subset of variables are selected based on

$$\widehat{M}^{\mathrm{CQR}} = \left\{ 1 \le j \le p : \left\| \widehat{D}_{nj}^{\mathrm{CQR}} \right\|_n^2 \ge \nu_n \right\},$$

where ν_n is a predefined threshold value. Since we are performing indepen-

dence screening, which means that each variable is applied independently as a predictor to decide its usefulness for predicting the response, the screening procedure can be implemented fast even in ultrahigh-dimensional space.

2.1.4 Weighted Composite Quantile Utility

Without considering the actual amounts of contributions from quantiles and the possible correlations between them, the CQR method endows different QR models with the same weight. Thus, more efficiency can be gained if appropriate weights are used by considering possible variation of contributions and their correlations between multiple quantiles, which leads to the weighted composite quantile regression (WCQR) estimator

$$\left(\left\{\hat{b}_j^{\mathrm{WCQR}}\left(\tau_k\right)\right\}_{k=1}^{K}, \hat{\beta}_j^{\mathrm{WCQR}}\right) = \arg\min_{b,\beta} \sum_{k=1}^{K} \varpi_k \sum_{i=1}^{n} \rho_{\tau_k}\left(Y_i - x_{ij}\beta - b\left(\tau_k\right)\right), \tag{2.8}$$

where $\boldsymbol{\varpi} = \left(\varpi_1, \ldots, \varpi_K\right)^{\mathrm{T}}$ is the vector of weights. In Theorem 5.2 (page 169) of Koenker [17], under certain conditions, as the sample size $n \to \infty$,

$$\sqrt{n}\left(\hat{\beta}_j^{\mathrm{WCQR}} - \beta\right) \to N\left(0, H\left(\varpi\right)\boldsymbol{\Sigma}_X^{-1}\right),$$

where $\boldsymbol{\Sigma}_{X_j}$ is the covariance of X_j, which is fixed and

$$H\left(\boldsymbol{\varpi}\right) = \frac{\boldsymbol{\varpi}^{\mathrm{T}} B \boldsymbol{\varpi}}{\boldsymbol{\varpi}^{\mathrm{T}}\mathbf{v}\mathbf{v}^{\mathrm{T}}\boldsymbol{\varpi}} \geq \left(\mathbf{v}^{\mathrm{T}} B \mathbf{v}\right)^{-1}.$$

The covariance matrix of $\hat{\beta}_j^{\mathrm{WCQR}}$ depends on the weights through $H\left(\varpi\right)$. Thus a natural way to select optimal weights in the WAQR is to minimizes $H\left(\varpi\right)$ in (2.9) to get the smallest variance of the estimator. The optimal weight $\boldsymbol{\varpi}_{\mathrm{opt}}$

is given by

$$\boldsymbol{\varpi}_{\text{opt}} = B^{-1}\mathbf{v}. \tag{2.9}$$

However, this optimal weight can be negative, which makes the minimization problem nonconvex. Therefore, we want to restrict the optimal weight to non-negative, namely, $\varpi_k \geq 0, k = 1, \ldots, K$, so that it lead to WCQR a optimal convex combination of quantile regression. Zhao and Xiao [33] derived the optimal weight in the following form $\boldsymbol{\varpi}^{\star}_{\text{opt}} = \left(\varpi^{\star}_{\text{opt},1}, \ldots, \varpi^{\star}_{\text{opt},K}\right)^{\text{T}}$

$$\varpi^{\star}_{\text{opt},k} = \left(2v_k - v_{k-1} - v_{k+1}\right) / \left(v_1 + v_K\right) \geq 0, \quad k = 1, \ldots, K. \tag{2.10}$$

For convenience we write $v_0 = v_{K+1} = 0$. However the WCQR with the constraint of non-negative weights needs additional assumption, that is, the log density function $\log f\left(u\right)$ of ε is concave.

If X_j and Y are independent, $D^{\text{WCQR}}_{nj} = \sum_{k=1}^{K} \varpi_k \left[Q_{\tau_k}\left(Y|X_j\right) - Q_{\tau_k}\left(Y\right)\right]$ is expected to be close to zero. Then an estimator of D^{WCQR}_{nj} is defined as follows:

$$\widehat{D}^{\text{WCQR}}_{nj} = \sum_{k=1}^{K} \varpi^{\star}_{\text{opt},k} \left(X_j \hat{\beta}^{\text{WCQR}}_j + \hat{b}^{\text{WCQR}}_j\left(\tau_k\right) - \widehat{F}^{-1}_{Y,n}\left(\tau_k\right)\right)$$

If X_j and Y are independent, $\widehat{D}^{\text{WCQR}}_{nj}$ is expected to be close to zero.

The independence screening is based on the magnitude of the estimated marginal components

$$\left\|\widehat{D}^{\text{WCQR}}_{nj}\right\|^2_n = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} \varpi^{\star}_{\text{opt},k} \left(x_{ij} \hat{\beta}^{\text{WCQR}}_j + \hat{b}^{\text{WCQR}}_j\left(\tau_k\right) - \widehat{F}^{-1}_{Y,n}\left(\tau_k\right)\right)\right]^2. \tag{2.11}$$

According to the value of $\left\|\widehat{D}_{nj}^{\mathrm{WCQR}}\right\|_n^2$, we can select this subset of variables

$$\widehat{M}_{\tau}^{\mathrm{WCQR}} = \left\{1 \leq j \leq p : \left\|\widehat{D}_{nj}^{\mathrm{WCQR}}\right\|_n^2 \geq \nu_n\right\},$$

where ν_n is a predefined threshold value. Because of the independence screening, the speed of screening is fast. In addition, as some of the weights in $\boldsymbol{\varpi}_{\mathrm{opt}}^{\star}$ are zero, it will make WCQR method computationally less intensive than CQR [2]. From our experience in large p and small n situations, this reduction tends to be significant.

The WAQR estimator differs from the WCQR estimator in several aspects. While the WCQR estimator in (2.8) is based on an aggregation of several quantile loss functions, the WAQR in (2.4) is based on a weighted average of separate estimators from different quantiles. As a result, computing the WAQR only involves K separate $(p+1)$-parameter minimization problems, whereas the WCQR requires solving a larger $(p+1)K$-parameter minimization problem. In addition, to ensure a proper loss function, the weights ϖ_k in WCQR are restricted to be non-negative; by contrast, the weights ω_k in WAQR can be negative. It is computationally appealing to impose less constraint on the weights. Zhao and Xiao [33] showed that under some conditions, the optimal WCQR is asymptotically equivalent to the optimal WAQR, and both are asymptotically efficient.

2.2 Nonlinear Model

In ultrahigh-dimensional data analysis, we usually know little about the form of the actual model. Therefore, besides the sparsity assumption, it is more

appropriate not to impose a specific model structure but allow the predictor effects to be nonlinear. To approximate the conditional quantiles, we use B-spline approximations, which are widely applied by, for example, Hall and Miller [10] as well as Fan, Feng and Song [4]. The choice of B-spline is very important to guarantee certain accuracy. However, in our setting, the accuracy is not our concern. We are only interested in picking out those variables related to the conditional quantiles. Therefore, only a few number of B-spline can capture the major information from the candidate predictors. The discussion of fine choice of B-spline is beyond the scope of this thesis. We follow the typical choice of Fan, Feng and Song [4] and select the number of B-spline basis functions as 5.

In this section, we study how to do variable screening in nonlinear case using our developed methods. The nonlinear case was discussed in He, Wang and Hong [11] for single quantile screening.

Without loss of generality, we assume that each $X_j \in [0,1], j = 1, \ldots, p$. Let $\mathbb{F}$ be the class of functions defined on $[0,1]$ whose lth derivative satisfies a Lipschitz condition of order c :

$$\left| f^{(l)}(s) - f^{(l)}(t) \right| \leqslant c_0 |s-t|^c,$$

for some positive constant $c_0, s, t \in [0,1]$, where l is a nonnegative integer and $c \in [0,1]$ satisfies $d = l + c > 0.5$. Let $0 = s_0 < s_1 < \ldots < s_m = 1$ be a partition of the interval. Using $s_i, i = 0, \ldots, m$ as knots, we construct $N = m + l$ normalized B-spline basis functions of order $l+1$ which form a basis for $\mathbb{F}$. We write these basis functions as a vector $\pi(t) = (B_1(t), \ldots, B_N(t))^{\mathrm{T}}$, where $\|B_m(\cdot)\|_\infty \leqslant 1$ and $\|\cdot\|_\infty$ denotes the sup norm.

Now we put $f_j(X_j) = Q_\tau(Y|X_j)$, $f_j(t) \in \mathbb{F}$. Then $f_j(t)$ can be well approximated by a linear combination of the basis functions $\pi(t)^{\mathrm{T}}\boldsymbol{\beta}$, for some $\boldsymbol{\beta} \in \mathbb{R}^N$ [11]. Let

$$\left(\hat{\boldsymbol{\beta}}_j(\tau), \hat{b}_j(\tau)\right) = \arg\min_{\boldsymbol{\beta},b} \sum_{i=1}^{n} \rho_\tau\left(Y_i - \pi(x_{ij})^{\mathrm{T}}\boldsymbol{\beta} - b\right). \tag{2.12}$$

As in linear model, this minimization problem can be solved by the interior point method [25], simplex method [19] or MM algorithm [13] easily with existing statistical software, like, R and SAS. Furthermore, we define

$$\hat{f}_{nj}(t) = \pi(t)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_j(\tau) + \hat{b}_j(\tau) - \widehat{F}^{-1}_{Y,n}(\tau),$$

where $\widehat{F}^{-1}_{Y,n}(\tau)$ is the τth sample quantile function based on $Y_1, \ldots, Y_n$. Thus $\hat{f}_{nj}(t)$ is a nonparametric estimator of $f_{nj}(t) = Q_\tau(Y|X_j) - Q_\tau(Y)$. According to (2.1), $\hat{f}_{nj}$ measures the difference between the conditional and unconditional quantile of Y and is expected to be close to zero if X_j is independent of Y.

The QaSIS is based on the magnitude of the estimated marginal components

$$\left\|\hat{f}_{nj}\right\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{nj}(x_{ij})^2.$$

As the final step, we select this subset of variables

$$\widehat{M}_\tau = \left\{1 \le j \le p : \left\|\hat{f}_{nj}\right\|_n^2 \ge \nu_n\right\},$$

where ν_n is a predefined threshold value. For the choice of possible threshold values, see Section 2.4.

When the conditional quantiles of the response have continuous, similar or

even exactly the same shape in certain quantile intervals, which is quite common in practice, the coefficients of the B-spline basis are continuous and even constant in those intervals. We assume the conditional quantiles are approximated by the same set of B-spline basis. The marginal quantile regression approach does not take such shared information across quantiles into account, which may lose efficiency. In this situation, we propose to combine information obtained from different quantiles. Suppose we want to consider K quantile levels, say $\{\tau_k \in (0,1), k = 1, \ldots, K\}$. Taking average of the estimated coefficients via AQR is a simple way to achieve that. Considering that the amounts of contribution made by different quantiles may be unequal, the optimal combination of the estimated coefficients obtained by WAQR gains more efficiency. When the coefficients across quantile levels are constant, CQR is employed to exactly capture this feature. Furthermore, optimally combining the QR loss functions using WCQR will increase the efficiency.

In the following subsections, we describe how we conduct variable screening using the four proposed estimators in nonlinear case. Once we choose the set of B-spline basis, the nonlinear screening will become exactly the same as the linear screening. What we need keep in mind is that for each predictor, we have more than one coefficients to estimate, as there is a set of B-spline basis related to them. The number of coefficients equals to the number of the basis. Except for that, the screening procedure keeps the same. For the completeness of the thesis, we briefly describe a little bit details on the nonlinear screening procedure.

2.2.1 Average Quantile Utility

Now in order to enhance the efficiency of the screening procedure for nonlinear models, we consider the AQR estimator. Consider multiple quantile levels $\{\tau_k \in (0,1), k = 1, \ldots, K\}$. For each τ_k, we estimate $\boldsymbol{\beta}$ via the marginal quantile regression and define the average quantile regression (AQR) estimator for nonlinear model as

$$\hat{\boldsymbol{\beta}}_j^{\mathrm{AQR}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\boldsymbol{\beta}}_j(\tau_k), \qquad \hat{b}_j^{\mathrm{AQR}} = \frac{1}{K}\sum_{k=1}^{K} \hat{b}_j(\tau_k),$$

where $\hat{\boldsymbol{\beta}}_j(\tau_k)$ and $\hat{b}_j(\tau_k)$ are defined in (2.12) at the quantile level τ_k. Note that both of them are vectors instead of scalars.

We define

$$\begin{aligned}\hat{f}_{nj}^{\mathrm{AQR}}(t) &= \pi(t)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j^{\mathrm{AQR}} + \hat{b}_j^{\mathrm{AQR}} - \frac{1}{K}\sum_{k=1}^{K} \widehat{F}_{Y,n}^{-1}(\tau) \\ &= \frac{1}{K}\sum_{k=1}^{K}\left[\pi(t)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j(\tau_k) + \hat{b}_j(\tau_k) - \widehat{F}_{Y,n}^{-1}(\tau_k)\right]\end{aligned}$$

as a nonparametric estimator of $f_{nj}^{\mathrm{AQR}} = K^{-1}\sum_{k=1}^{K}[Q_{\tau_k}(Y|X_j) - Q_{\tau_k}(Y)]$. $\hat{f}_{nj}^{\mathrm{AQR}}$ is expected to be around zero if X_j is not dependent of Y.

Actually this independence screening procedure is based on the magnitude of the estimated marginal components, which is given by

$$\begin{aligned}\left\|\hat{f}_{nj}^{\mathrm{AQR}}\right\|_n^2 &= \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{nj}^{\mathrm{AQR}}(x_{ij})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\pi(x_{ij})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j(\tau_k) + \hat{b}_j(\tau_k) - \widehat{F}_{Y,n}^{-1}(\tau_k)\right)\right)^2.\end{aligned}$$

Then the subset of variables is selected as follows:

$$\widehat{M}^{\mathrm{AQR}} = \left\{ 1 \leq j \leq p : \left\| \hat{f}_{nj}^{\mathrm{AQR}} \right\|_n^2 \geq \nu_n \right\},$$

where ν_n is a predefined threshold value.

Because of the independence screening, that is, each variable is used independently as a predictor to decide its usefulness for predicting the response, the screening can be done very fast even in the ultrahigh-dimensional case.

2.2.2 Weighted Average Quantile Utility

Since the AQR method is equally weighted for different quantile levels. Therefore, it may be more effective if appropriate weights are applied by using the WAQR estimator:

$$\hat{\boldsymbol{\beta}}^{\mathrm{WAQR}} = \sum_{k=1}^{K} \omega_k \hat{\boldsymbol{\beta}} \left(\tau_k \right), \qquad \hat{b}^{\mathrm{WAQR}} = \sum_{k=1}^{K} \omega_k \hat{b} \left(\tau_k \right), \tag{2.13}$$

where $\hat{\boldsymbol{\beta}}_j \left(\tau_k \right)$ and $\hat{b}_j \left(\tau_k \right)$ are defined in (2.12) at the quantile level τ_k, $\boldsymbol{\omega} = \left(\omega_1, \ldots, \omega_K \right)^{\mathrm{T}}$ is a vector of weights. In AQR, all the weights are $1/K$. For WAQR, we can choose an optimal weighting scheme to obtain the most efficiency. The optimal weight $\boldsymbol{\omega}_{\mathrm{opt}} = \left(\omega_{\mathrm{opt},1}, \ldots, \omega_{\mathrm{opt},K} \right)^{\mathrm{T}}$ is the same as that in the linear situation (see (2.6)).

We define

$$\begin{aligned} \hat{f}_{nj}^{\mathrm{WAQR}} \left(t \right) &= \pi \left(t \right)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j^{\mathrm{WAQR}} + \hat{b}_j^{\mathrm{WAQR}} - \sum_{k=1}^{K} \omega_{\mathrm{opt},k} \left(\widehat{F}_{Y,n}^{-1} \left(\tau_k \right) \right) \\ &= \sum_{k=1}^{K} \omega_{\mathrm{opt},k} \left[\pi \left(t \right)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j \left(\tau_k \right) + \hat{b}_j \left(\tau_k \right) - \widehat{F}_{Y,n}^{-1} \left(\tau_k \right) \right] \end{aligned}$$

as the nonparametric estimator of $f_{nj}^{\mathrm{WAQR}} = \sum_{k=1}^{K} \omega_k \left[Q_{\tau_k}\left(Y|X_j\right) - Q_{\tau_k}\left(Y\right) \right]$. As mentioned before, if X_j is an important variable, then $\hat{f}_{nj}^{\mathrm{WAQR}}$ is supposed to be away from zero, as it can capture the effect that X_j brings in to change the conditional and unconditional quantiles of Y.

We rank and screen the variables based on the magnitude of the estimated marginal components

$$\begin{aligned}\left\| \hat{f}_{nj}^{\mathrm{WAQR}} \right\|_n^2 &= \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{nj}^{\mathrm{WAQR}} \left(x_{ij}\right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \omega_{\mathrm{opt},k} \left(\pi\left(x_{ij}\right)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j\left(\tau_k\right) + \hat{b}_j\left(\tau_k\right) - F_{Y,n}^{-1}\left(\tau_k\right) \right) \right)^2,\end{aligned}$$

and focus on the following subset of variables

$$\widehat{M}^{\mathrm{WAQR}} = \left\{ 1 \leq j \leq p : \left\| \hat{f}_{nj}^{\mathrm{WAQR}} \right\|_n^2 \geq \nu_n \right\},$$

where ν_n is a predefined threshold value.

2.2.3 Composite Quantile Utility

When the coefficients are constant, we propose to combine the information from multiple quantile levels using CQR. We estimate the model over multiple quantile levels $\left\{ \tau_k \in (0,1), k = 1, \ldots, K \right\}$ jointly based on the following modified quantile loss function:

$$\left(\left\{ \hat{b}_j^{\mathrm{CQR}}\left(\tau_k\right) \right\}_{k=1}^{K}, \hat{\boldsymbol{\beta}}_j^{\mathrm{CQR}} \right) = \underset{b, \boldsymbol{\beta}}{\arg\min} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k} \left(Y_i - \pi\left(x_{ij}\right)^{\mathrm{T}} \boldsymbol{\beta} - b\left(\tau_k\right) \right).$$

Typically, we use the equally spaced quantiles: $\tau_k = \frac{k}{K+1}$ for $k = 1, \ldots, K$. We solve it using the MM algorithm in R as well.

Furthermore, we define

$$\hat{f}_{nj}^{\mathrm{CQR}}(t) = \sum_{k=1}^{K} \left[\pi(t)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j^{\mathrm{CQR}} + \hat{b}_j^{\mathrm{CQR}}(\tau_k) - \widehat{F}_{Y,n}^{-1}(\tau_k) \right]$$

as an nonparametric estimator of $f_{nj}^{\mathrm{CQR}} = \sum_{k=1}^{K} [Q_{\tau_k}(Y|X_j) - Q_{\tau_k}(Y)]$. As the equivalent relationship (2.1) holds for any given $\tau \in (0,1)$, $\hat{f}_{nj}^{\mathrm{CQR}}$ is expected be close to zero if X_j and Y are independent.

Next, we rank the features by the magnitude of the estimated marginal components

$$\begin{aligned} \left\| \hat{f}_{nj}^{\mathrm{CQR}} \right\|_n^2 &= \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{nj}^{\mathrm{CQR}}(x_{ij})^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \left(\pi(x_{ij})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j^{\mathrm{CQR}} + \hat{b}_j^{\mathrm{CQR}}(\tau_k) - \widehat{F}_{Y,n}^{-1}(\tau_k) \right) \right]^2 . \end{aligned}$$

In the end, based on the value of $\left\| \hat{f}_{nj}^{\mathrm{CQR}} \right\|_n^2$, we select the subset of variables

$$\widehat{M}^{\mathrm{CQR}} = \left\{ 1 \le j \le p : \left\| \hat{f}_{nj}^{\mathrm{CQR}} \right\|_n^2 \ge \nu_n \right\},$$

where ν_n is a predefined threshold value.

2.2.4 Weighted Composite Quantile Utility

Considering the actual amounts of contributions from quantiles and the possible correlations between them, more efficiency can be gained if appropriate

weights are used by WCQR

$$\left(\left\{\hat{b}_j^{\mathrm{WCQR}}\left(\tau_k\right)\right\}_{k=1}^{K}, \hat{\boldsymbol{\beta}}_j^{\mathrm{WCQR}}\right) = \arg\min_{b,\boldsymbol{\beta}} \sum_{k=1}^{K} \varpi_k \sum_{i=1}^{n} \rho_{\tau_k}\left(Y_i - \pi\left(x_{ij}\right)^{\mathrm{T}} \boldsymbol{\beta} - b\left(\tau_k\right)\right),$$

where $\boldsymbol{\varpi} = \left(\varpi_1, \ldots, \varpi_K\right)^{\mathrm{T}}$ is the vector of weights.

If X_j and Y are independent, $f_{nj}^{\mathrm{WCQR}} = \sum_{k=1}^{K} \varpi_k \left[Q_{\tau_k}\left(Y|X_j\right) - Q_{\tau_k}\left(Y\right)\right]$ is expected to be close to zero. Then an nonparametric estimator of f_{nj}^{WCQR} is defined as follows:

$$\hat{f}_{nj}^{\mathrm{WCQR}}\left(t\right) = \sum_{k=1}^{K} \varpi_{\mathrm{opt},k}^{\star}\left[\pi\left(t\right)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j^{\mathrm{WCQR}} + \hat{b}_j^{\mathrm{WCQR}}\left(\tau_k\right) - \widehat{F}_{Y,n}^{-1}\left(\tau_k\right)\right]$$

where $\boldsymbol{\varpi}_{\mathrm{opt}}^{\star} = \left(\varpi_{\mathrm{opt},1}^{\star}, \ldots, \varpi_{\mathrm{opt},K}^{\star}\right)^{\mathrm{T}}$ is the optimal weight in (2.10). If X_j is not an important variable, $\hat{f}_{nj}^{\mathrm{WCQR}}\left(t\right)$ is expected to be close to zero.

Then we also rank the variables by the magnitude of the estimated marginal components

$$\begin{aligned}\left\|\hat{f}_{nj}^{\mathrm{WCQR}}\right\|_n^2 &= \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{nj}^{\mathrm{WCQR}}\left(x_{ij}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} \boldsymbol{\varpi}_{\mathrm{opt},k}^{\star}\left(\pi\left(x_{ij}\right)^{T} \hat{\boldsymbol{\beta}}_j^{\mathrm{WCQR}} + \hat{b}_j^{\mathrm{WCQR}}\left(\tau_k\right) - \widehat{F}_{Y,n}^{-1}\left(\tau_k\right)\right)\right]^2.\end{aligned}$$

Then we will select the subset of variables

$$\widehat{M}^{\mathrm{WCQR}} = \left\{1 \leq j \leq p : \left\|\hat{f}_{nj}^{\mathrm{WCQR}}\right\|_n^2 \geq \nu_n\right\},$$

where ν_n is a predefined threshold value.

2.3 Estimating $f(Q(\tau))$

The way to find optimal weights for WAQR and WCQR is to minimize the covariance matrices of their estimators to achieve the maximum efficiency. Both optimal weights have an unknown term $f(Q(\tau))$. In order to compute the optimal weights, we need to estimate it. In practice it is not easy to estimate $f(Q(\tau))$ properly. There are a lot of literatures on how to estimate $f(Q(\tau))$, for example, Koenker [17] provided a method. Let

$$s(\tau) = [f(Q(\tau))]^{-1}.$$

Differentiating the identity $F(Q(t)) = t$, we find that $s(t)$ is simply the derivative of the quantile function; that is

$$\frac{d}{dt}Q(t) = s(t).$$

It is therefore natural to estimate $s(t)$ by using a simple difference quotient of the empirical quantile function:

$$\hat{s}_n(t) = \frac{\widehat{Q}_n(t+h_n) - \widehat{Q}_n(t-h_n)}{2h_n},$$

where $\widehat{Q}$ is an estimate of Q and h_n is a bandwidth that tends to zero as $n \to \infty$. Provided that the τth conditional quantile function of Y is linear which is one of the cases in this thesis, then for $h_n \to 0$ we can consistently estimate the parameters of the $\tau \pm h_n$ conditional quantile function by $\hat{\boldsymbol{\beta}}(\tau \pm h_n)$. And

the density $f_i(Q_i)$ can thus be estimated by the difference quotient

$$\hat{f}_i(Q_i(\tau)) = \frac{2h_n}{\mathbf{x}_i^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}}(t+h_n) - \hat{\boldsymbol{\beta}}(t-h_n)\right)}.$$

A potential difficulty with the proposed estimator $f_i(Q_i(\tau))$ is that there is no guarantee of positivity for every observation in the sample. In the implementation of this approach, we simply replace $\hat{f}_i(Q_i(\tau))$ by its positive part, that is,

$$\hat{f}_i^{+}(Q_i(\tau)) = \max\left(0, \frac{2h_n}{\mathbf{x}_i^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}}(t+h_n) - \hat{\boldsymbol{\beta}}(t-h_n)\right) - \epsilon}\right),$$

where $\epsilon > 0$ is a small tolerance parameter intended to avoid dividing by zero in the (rare) case in which $\mathbf{x}_i^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}}(t+h_n) - \hat{\boldsymbol{\beta}}(t-h_n)\right) = 0$. This method can be easily extended to our approach. For simplicity, when we estimate the optimal weights for WAQR and WCQR, $\hat{\boldsymbol{\beta}}$ can be the estimators of AQR and CQR respectively. If the τth conditional quantile function of Y is nonlinear, then B-spline approximations can also be used.

The aforementioned method works reasonably well. However, in this thesis, we adopt another competing method used by Zhao and Xiao [33], which is simpler:

(i) Apply uniform weights $1/K$ to obtain the preliminary estimator $\hat{\boldsymbol{\beta}}$, and compute the "residuals" as $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$.

(ii) Use the nonparametric kernel density estimator to estimate $f(u)$:

$$\hat{f}(u) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{u - \hat{\varepsilon}_i}{b}\right), \tag{2.14}$$

where $K(\cdot)$ is a non-negative kernel function and $b > 0$ is a bandwidth.

(iii) Estimate $f(Q(\tau))$ by $\hat{f}\left(\hat{Q}(\tau)\right)$, where $\hat{Q}(\tau)$ is the τth sample quantile of $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$.

In (2.14), we use Gaussian kernel for K and choose

$$b = 0.9 * \min\left\{\mathrm{sd}\left(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\right), \frac{\mathrm{IQR}\left(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\right)}{1.34}\right\} * n^{-1/5},$$

where "sd" and "IQR" denote the sample standard deviation and sample interquartile respectively.

2.4 Threshold Rule

Under some conditions, as $n \to \infty$, the magnitude of the estimators of the difference between the unconditional and conditional quantiles of Y always ranks an active predictor above an inactive one in probability. Guaranteeing a clear separation between the active and inactive predictors is very important [34]. An inappropriate threshold value may provide inaccurate information, as active predictors may be missed or inactive predictors may be selected. There are a few papers that discuss how to choose the threshold, for example, Zhu et al. [34]. Because of the importance of threshold, it deserves more discussion, especially in our new proposed methods. In this section, we propose several threshold rules to obtain a cutoff value to separate the active and inactive predictors.

Generally speaking, there are two different threshold rules: soft and hard threshold rules. In the soft threshold rule, artificial auxiliary variables are added to the data, which was first proposed by Luo, Stefanski and Boos [24]

and then extended by Wu, Boos and Stefanski [30]. We adopt the same idea in our setup as used by Zhu et al. [34]. We now take the linear model as an example to explain how the soft threshold rule works. We independently and randomly generate d auxiliary variables $\mathbf{Z} \sim N_d(\mathbf{0}, \mathbf{I}_d)$ such that $\mathbf{Z}$ is independent of both $\mathbf{X}$ and Y. The normality assumption is not critical here and other distributions can be chosen as well. Given a random sample $\{\mathbf{z}_i, i = 1, \ldots, n\}$, we treat $(\mathbf{X}, \mathbf{Z})$ as the predictors and Y as the response. Since $\mathbf{Z}$ is truly inactive by construction, as $n \to \infty$, $\|\widehat{D}_{nj}\|_n^2$ always ranks an active predictor above an inactive one in probability, where $\|\widehat{D}_{nj}\|_n^2$ represents the magnitude of the estimators of the difference between unconditional and conditional quantiles of Y for all the cases in linear models for brevity. Mathematically, it holds in probability that

$$\min_{j \in M} \left\| \widehat{D}_{nj} \right\|_n^2 > \max_{l=1,\ldots,d} \left\| \widehat{D}_{n(p+l)} \right\|_n^2$$

by Theorem 2 of Zhu et al. [34], where M is the true model.

Now we define $C_d = \max_{l=1,\ldots,d} \|\widehat{D}_{n(p+l)}\|_n^2$, which can be viewed as a benchmark that separates the active predictors from the inactive ones. In this case, we get the following selection rule:

$$\widehat{M}^1 = \left\{ j : \left\| \widehat{D}_{nj} \right\|_n^2 > C_d \right\}. \tag{2.15}$$

We call it the soft threshold selection.

An issue of practical interest in soft threshold is the choice of the number of auxiliary variables d. Intuitively, a small d value may introduce much variability, whereas a large d value requires heavy computation. Empirically, we

choose $d = p$ [34].

In addition to soft threshold, we also consider a hard threshold rule proposed by Fan and Lv [7], which retains a fixed number of predictors with the largest N values of $\|\hat{D}_{nj}\|_n^2$'s. Mathematically, the hard threshold rule can be expressed as follows:

$$\widehat{M}^2 = \left\{ j : \left\| \widehat{D}_{nj} \right\|_n^2 > \left\| \widehat{D}_{n(N)} \right\|_n^2 \right\}, \tag{2.16}$$

where N is usually chosen to be $[n/\log n]$ and $\|\widehat{D}_{n(N)}\|_n^2$ denotes the Nth largest value among all $\|\widehat{D}_{nj}\|_n^2$'s. Usually N is large enough to guarantee that all the important variables are kept. We then use some well-developed variable selection methods, such as the adaptive LASSO, SCAD and so on to remove those irrelevant variables. Hard threshold is simple and does not need extra computational burden. Therefore, it is also very popular in practice.

In practice, the data determine whether the soft or hard threshold comes into play. In order to take advantage of both methods and avoid missing important variables and selecting irrelevant variables, we propose to combine the soft and hard threshold as in Zhu et al. [34], and construct the final active predictor index set as

$$\widehat{M} = \widehat{M}^1 \cup \widehat{M}^2. \tag{2.17}$$

To better understand the two threshold rules, we conducted a simulation study. We make the following observations from our simulation study. When the number of active variable is small, the hard threshold rule often dominates the soft selection rule. On the other hand, when there are many active predic-

tors, the soft threshold becomes more dominant. While the hard threshold is fully determined by the sample size, soft threshold takes into account the effect of signals in the data, which is helpful when the number of active predictors is relatively large.

Chapter 3

Numerical Studies

3.1 Monte Carlo Studies

3.1.1 General Setup

In this section we assess the finite sample performance of the proposed methods and compare them with other approaches via Monte Carlo studies. For brevity, we denote our four combining quantile adaptive sure independence screening approaches as average quantile regression screening (AQRS), weighted average quantile regression screening (WAQRS), composite quantile regression screening (CQRS) and weighted composite quantile regression screening (WCQRS).

Except for Example 1, we consider six different distributions for the error term $\boldsymbol{\varepsilon}$ in all the other examples:

1. the standard normal distribution;

2. student-t distribution with one degree of freedom, that is, Cauchy distribution;

3. the standard normal with 10% outliers following student-t distribution with one degree of freedom;

4. the standard normal with 20% outliers following student-t distribution with one degree of freedom;

5. normal mixture distribution 1: $0.9N(0,1) + 0.1N(10,1)$;

6. normal mixture distribution 2: $0.8N(0,1) + 0.2N(10,1)$.

The SIS performs well under normal distribution, so we take it into account to compare our proposed screening methods with SIS. t_1 distribution, which is a heavy-tailed distribution, has infinite variance. The third and fourth distributions have variance outliers. The last two have location outliers. Note that only if the log density function $\log f(u)$ of ε is concave, the positive optimal weight of WCQR can be obtained. Due to the constraint of that condition, we do not consider the second, fifth and sixth cases of the error distributions in WCQRS method. Throughout, we consider combining information over 9 quantiles $\tau_k = k/10, k = 1, \ldots, 9$. For each scenario, we repeat 100 times. To compute QaSIS procedure of He, Wang and Hong [11], NIS procedure of Fan, Feng and Song [4] as well as our procedure in nonlinear models, the number of B-spline basis functions is set to be 5.

To evaluate the performance of our screening methods, we consider two criteria as Zhu et al. [34]. The first criterion is the minimum model size, that is the smallest number of predictors that we need to include to ensure that all the active variables are selected. We denote this number by $\mathcal{R}$. Note that the first criterion does not need to specify a threshold. Based on the result of Example 1, we propose to use hard threshold rule in the other examples

to reduce the computational burden. The second criterion focuses on the proportion of active variables selected by the screening procedure with the threshold $\nu_n = [n/\log(n)]$. We denote this proportion by $\mathcal{S}$. Since screening usually serves as a preliminary massive reduction step, and is often followed by a variable selection, it is important to retain all the active variables. A competent variable screening procedure is expected to have the value of $\mathcal{R}$ reasonably small comparing to the number of active variables and the value of $\mathcal{S}$ close to one. For each example, we record the median and interquartile range (IQR) of $\mathcal{R}$ and $\mathcal{S}$ respectively, as well as the median of running time T. For QaSIS, we give the results for two quantiles $\tau = 0.5$ and $\tau = 0.75$.

The QR, AQR, WAQR estimators are obtained by using *quantreg* package in R and the R codes for CQR and WCQR estimators are adapted from the MATLAB code written by Kai, Li and Zou [15] using MM algorithm. For more details on MM algorithm, see Hunter and Lange [13]. Due to the high dimensionality, all the simulation and real data examples are run on the cluster of WestGrid.

3.1.2 Threshold Rule

Appropriately choosing threshold value can be able to separate the active and inactive predictors. To better understand the proposed threshold rules and see how the data determine whether the soft or hard threshold comes into play, we apply AQRS method to the linear model as illustration:

Example 1 $(n = 200, p = 2000)$. Consider the linear model (2.2)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the predictors $\mathbf{X}$ are generated from the multivariate normal distribution with mean $\mathbf{0}$ and the covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{2000\times 2000}$ with $\sigma_{ii} = 1$, $\sigma_{ij} = 0.4$ if $i, j \in M$ or $i, j \in M^C$, and $\sigma_{ij} = 0.1$ otherwise. The error $\boldsymbol{\varepsilon}$ follows a standard normal distribution. $\boldsymbol{\beta} = (2 - U_1, \ldots, 2 - U_s, 0, \ldots, 0)^{\mathrm{T}}$, and U_k's follows a uniform distribution on $[0, 1]$.

We report the total number of selected predictors determined by the soft selection rule 2.15, the hard threshold rule (2.15) and the combination of two threshold rules (2.17) in Table (3.2), and record the minimum, the first quartile, the median, the third quartile and the maximum of the selected size $\mathcal{R}$ in 100 data replications, with an increasing number of truly active predictors $s = 4, 8, 16, 32, 64, 96$. The maximum number of selected predictors is set no greater than n. For hard threshold rule, we use $[n/\log(n)] = 37$.

As a result, we make roughly the same observations as Li et al. [34]. When the signal in the data is sparse (a small s), the hard threshold rule tends to dominate the selection rule. This is reflected by the simulation result that the total number of selected predictors often equals 37 when $s = 4, 8$ and 16. On the other hand, when there are many active predictors (a large s), the soft threshold becomes more dominant. While hard threshold is fully determined by the sample size, soft threshold takes into account the effect of signals in the data, which is especially helpful when s is relatively large.

3.1.3 Linear Models

We begin with a class of linear models and compare the performance of AQRS, WAQRS, CQRS and WCQRS with SIS and QaSIS.

Example 2 $(n = 200, p = 1000)$. This example is adopted from the pre-

vious work by Fan and Lv [7]. We use the linear model (2.2)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the vector of covariates $\mathbf{X} = (X_1, \ldots, X_{1000})$ is generated from the standard normal distribution. The size s of the true model, i.e. the number of non-zero coefficients is chosen to be 8, and the non-zero components of the p-vectors $\boldsymbol{\beta}$ are randomly chosen as follows. We set $a = 4\log(n)/n^{1/2}$ and pick non-zero coefficients of the form $(-1)^u (a + |z|)$ for each model, where u is drawn from Bernoulli distribution with parameter 0.4 and z is drawn from the standard normal distribution.

Example 3 $(n = 200, p = 1000)$. In this example, we use similar model to that in Example 1, except that the predictors are now correlated with each other. To introduce correlation between predictors, we first use MATLAB function *sprandsym* to generate randomly an $s \times s$ symmetric positive definite matrix $\mathbf{A}$ with condition number $n^{1/2}/\log(n)$ and draw samples of s predictors $X_1, \ldots, X_s$ from $N(\mathbf{0}, \mathbf{A})$. Then we take $Z_{s+1}, \ldots, Z_p \sim N(\mathbf{0}, \mathbf{I}_{p-s})$ and define the remaining covariates as follows:

$$\begin{cases} X_i = Z_i + rX_{i-s}, & i = s+1, \ldots, 2s \\ X_i = Z_i + (1-r)X_1, & i = 2s+1, \ldots, p \end{cases}$$

with $r = 1 - 4\log(n)/p$.

Example 4 $(n = 200, p = 1000)$. This example is adapted from He, Wang and Hong. [11]. The random data are generated from

$$\mathbf{y} = X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5 + (X_{20} + X_{21} + X_{22}) \cdot \boldsymbol{\varepsilon}$$

where $\mathbf{X} = (X_1, \ldots, X_{1000})$ follows the multivariate normal distribution with mean $\mathbf{0}$ and the covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{1000\times 1000}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho^{|i-j|}$, $i \neq j$. Here we consider $\rho = 0.8$. Different from the regression models in Example 2 and 3, this model is heteroscedastic: the number of active variables is 5 at the median but 8 elsewhere.

We observe the following from the result of Example 2: (i) For normal distribution, SIS performs the best, and WAQRS is comparable with SIS. Our proposed four methods all outperform QaSIS where single quantile level is used. (ii) For t_1 distribution, SIS performs the worst, as t_1 has heavy tail. In this situation, our methods based on multiple quantile regression perform really well. However, Since QaSIS just takes an individual quantile level into consideration, which leads to miss some information from other parts, it performs not well. (iii) For the third and fourth distribution, because of the heavy-tailed outliers, SIS performs not well, while WAQRS exhibits the best performance. The other three proposed methods are comparable with WAQRS. (iv) For the two mixtures of normal distributions, SIS outperform QaSIS, while our four proposed methods perform better than SIS. (v) For all the six distribution, the weighted ones are comparable with or even outperform the average ones.

The following observation are from the result of Example 3: (i) For normal distribution, SIS performs the best, and WAQRS is comparable with SIS. Our proposed four methods all outperform QaSIS where single quantile level is used. (ii) For t_1 distribution, because of the heavy tail, SIS performs really bad, while CQRS performs pretty well. As QaSIS just considers only one quantile level, it performs not well. (iii) For the third distribution with small quantity of outliers, Our proposed methods are comparable with normal dis-

tribution. When the amount of outliers increases, SIS performs worse than our proposed four methods but still better than QaSIS. (iv) For the mixtures of two normal distributions, our four proposed methods are comparable with or even outperform SIS. However, SIS still exhibits better performance than QaSIS.

The following are observed from the result of Example 4: (i) Since it is a heteroscedastic model: the number of active variables is 5 at the median but 8 elsewhere, we can see SIS performs well only for normal distribution while it performs really bad for non-normal distributions. (ii) QaSIS performs pretty well at the median for all the distributions. (iii) For normal distribution, our proposed methods are all comparable with or even outperform SIS. (iv) For t_1 distribution, WAQRS and CQRS perform the best. (v) For the third and fourth distribution, AQRS, WAQRS, CQRS and WCQRS can almost achieve the same best performance, which are better than QaSIS at 0.75 quantile. (vi) Comparing the fifth and the sixth distribution, we can see that QaSIS performs worse with more location outliers. Our proposed methods still perform better than SIS and QaSIS.

3.1.4 Nonlinear Models

In this subsection, we try to demonstrate that our approaches are useful in the sense that they work well for a large variety of different models when there is little knowledge about the underlying true model. In order to support this assertion, we compare our approaches with NIS and QaSIS.

Example 5 $(n = 400, p = 1000)$. This example is adopted from He, Wang

and Hong [11]. First, we define the following functions

$$
\begin{aligned}
g_1(x) &= x; \\
g_2(x) &= (2x-1)^2; \\
g_3(x) &= \sin(2\pi x)/\left(2-\sin(2\pi x)\right); \\
g_4(x) &= 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 \\
&\quad + 0.5\sin(2\pi x)^3.
\end{aligned}
$$

Then the random data are generated from

$$
\mathbf{y} = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\boldsymbol{\varepsilon},
$$

where $\mathbf{X} = (X_1, \ldots, X_{1000})$ follows the multivariate normal distribution with the same correlation structure described in Example 4. We vary the value of ρ to be 0 and 0.6.

From the results of Example 5, we can see that (i) When $\rho = 0$, no method works well in terms of the minimum model size. This is because the independent signals work against the marginal effect estimation as accumulated noise, thus masking the relatively weak signals from X_3 and X_4 in this model. (ii) When $\rho = 0.6$, NIS has the worst performance for all the distribution including normal distribution. For the other five distributions, one of our proposed method exhibits the best performance and QaSIS can be comparable with some of our proposed method. For example, in the third distribution, QaSIS at median and WCQRS almost have the same performance.

Based on the results from Example 2 to Example 5, we have the following conclusion: (i) SIS and NIS procedure exhibit the best performance when

the random error has a normal distribution, but their performance deteriorate substantially for heavy-tailed or heteroscedastic errors. (ii) AQRS, WAQRS, CQRIS and WCQRS significantly outperform QaSIS which considers only a single quantile level. (iii) For almost all the cases discussed, weighted methods outperform equally weighted ones. (iv) For non-normal distributions, our four approaches substantially perform better than SIS and NIS, whereas they are comparable for $N(0,1)$. (v) We observe that in Example 5 when ρ =0, no method works well in terms of the minimum model size. This is because the independent signals work against the marginal effect estimation as accumulated noise, thus masking the relatively weak signals from X_3 and X_4 in this model. (vi) As for the running time T, SIS always runs the fastest due to the simple correlation. Our methods based on the four estimators are slower than the QaSIS because of the multiple quantile levels. The weighted one always runs slower than the average one since it takes time to estimate the optimal weights.

3.2 Real Data Analysis

As an illustration, we apply the proposed screening methods on the diffuse large-B-cell lymphoma (DLBCL) microarray data of Rosenwald et al. ([26]).

DLBCL is the most common type of lymphoma in adults and has only about 35 to 40 percent survival rate after the standard chemotherapy. Hence, it is meaningful to explore the genetic factors that influence the survival time. The data set contains the survival times of $n = 240$ patients and the gene expression measurements of $p = 7399$ genes for each patient. Given such a large number of predictors and small sample size, variable screening seems to

be a necessary initial step as a prelude to any other sophisticated statistical modeling that does not cope well with such high dimensionality.

All predictors, namely, the gene expression measurements for each gene, are standardized to have mean zero and variance one. We form the bivariate response consisting of the observed survival time and the censoring indicator. To assess the predictive performance of the proposed methods, we apply the data split by Bair and Tibshirani ([1]), which divides the data into a training set with $n_1 = 160$ patients and a testing set with remaining $n_2 = 80$ patients. The index of the training set is available at `http://www-stat.stanford.edu/~tibs/superpc/staudt.html`. First we apply the proposed screening methods to the training set. Since only a small number of genes are relevant, and according to the result of Example 1, the hard threshold is more dominant in this scenario. Therefore, we choose $[n_1/\log(n_1)] = 31$ genes in our final set. We evaluate the prediction performance of this model following the approach of Li and Luan ([21]) as well as Lu and Li ([23]). Specifically, we apply the screening approach and fit a Cox proportion hazards model for the training data. Then we compute the risk scores for the testing data and divide it into a low-risk group and a high-risk group, where the cutoff value is determined by the median of the estimated scores from the training set.

We note that choosing the Cox model after screening only serves as a simple illustration in this example. After variable screening, more refined model building and selection can be employed.

Figure (3.1) shows the Kaplan-Meier estimates of survival curves for both risk groups of patients in the testing data, as well as the p-values yielded from the log-rank test for each screening approach. The two curves of our AQRS, CQRS and WCQRS approaches are well separated, while the two curves of

NIS and QaSIS are less well separated. By comparing each p-value, we can see AQRS, CQRS and WCQRS approaches all outperform NIS and QaSIS. In addition, the p-value of WCQRS is smaller than that of CQRS, which indicates a better prediction of the fitted model. However, we can see that WAQRS does not perform well, even worse than AQRS. An acceptable explanation is that we fail to estimate the $f\left(Q(\tau)\right)$ properly and thus obtain the inappropriate optimal weights.

Table 3.1: Example 1: Threshold Rule

s		soft	hard	combination
4	$\mathcal{R}_{\min}$	4	37	37
	$\mathcal{R}_{0.25}$	6	37	37
	$\mathcal{R}_{0.5}$	9	37	37
	$\mathcal{R}_{0.75}$	30	37	37
	$\mathcal{R}_{\max}$	200	37	200
8	$\mathcal{R}_{\min}$	8	37	37
	$\mathcal{R}_{0.25}$	10	37	37
	$\mathcal{R}_{0.5}$	20	37	37
	$\mathcal{R}_{0.75}$	43	37	43
	$\mathcal{R}_{\max}$	200	37	200
16	$\mathcal{R}_{\min}$	16	37	37
	$\mathcal{R}_{0.25}$	18	37	37
	$\mathcal{R}_{0.5}$	23	37	37
	$\mathcal{R}_{0.75}$	40	37	40
	$\mathcal{R}_{\max}$	200	37	200
32	$\mathcal{R}_{\min}$	32	37	37
	$\mathcal{R}_{0.25}$	35	37	37
	$\mathcal{R}_{0.5}$	44	37	44
	$\mathcal{R}_{0.75}$	69	37	69
	$\mathcal{R}_{\max}$	200	37	200
64	$\mathcal{R}_{\min}$	64	37	64
	$\mathcal{R}_{0.25}$	69	37	69
	$\mathcal{R}_{0.5}$	79	37	79
	$\mathcal{R}_{0.75}$	119	37	119
	$\mathcal{R}_{\max}$	200	37	200
96	$\mathcal{R}_{\min}$	96	37	96
	$\mathcal{R}_{0.25}$	99	37	99
	$\mathcal{R}_{0.5}$	107	37	107
	$\mathcal{R}_{0.75}$	138	37	138
	$\mathcal{R}_{\max}$	200	37	200

Table 3.2: Example 2: Independent Model

$\boldsymbol{\varepsilon}$	Method	$\mathcal{R}_{0.5}$	$\mathcal{R}_{IQR}$	$\mathcal{S}_{0.5}$	$\mathcal{S}_{IQR}$	$T_{0.5}$
$N(0,1)$	SIS	13	19	1.00	0.00	1.09
	$\text{QaSIS}_{0.5}$	39	89	0.875	0.125	2.90
	$\text{QaSIS}_{0.75}$	63	156	0.875	0.25	2.69
	AQRS	16	29	1.00	0.09	16.00
	WAQRS	13	20	1.00	0.00	44.89
	CQRS	21	54	1.00	0.125	691.273
	WCQRS	20	45	1.00	0.125	884.21
t_1	SIS	445	393	0.625	0.25	0.99
	$\text{QaSIS}_{0.5}$	96	192	0.875	0.094	2.15
	$\text{QaSIS}_{0.75}$	208	292	0.75	0.125	2.05
	AQRS	45	115	0.875	0.125	15.73
	WAQRS	37	75	1.00	0.125	77.18
	CQRS	64	155	1.00	0.125	576.81
$0.9N(0,1)+0.1t_1$	SIS	40	67	0.875	0.125	1.00
	$\text{QaSIS}_{0.5}$	28	92	1.00	0.125	2.58
	$\text{QaSIS}_{0.75}$	73	140	0.875	0.25	2.62
	AQRS	18	33	1.00	0.125	21.45
	WAQRS	14	24	1.00	0.00	56.41
	CQRS	22	45	1.00	0.125	505.874
	WCQRS	21	81	1.00	0.125	1031.365
$0.8N(0,1)+0.2t_1$	SIS	102	152	0.875	0.125	1.04
	$\text{QaSIS}_{0.5}$	46	87	0.875	0.125	2.18
	$\text{QaSIS}_{0.75}$	97	167	0.875	0.125	2.07
	AQRS	22	52	1.00	0.125	21.23
	WAQRIS	15	40	1.00	0.125	53.32
	CQRS	26	37	1.00	0.125	490.866
	WCQRS	24	72	1.00	0.125	1171.867
$0.9N(0,1)+0.1N(10,1)$	SIS	30	51	1.00	0.125	1.20
	$\text{QaSIS}_{0.5}$	52	83	0.875	0.125	2.96
	$\text{QaSIS}_{0.75}$	143	190	0.875	0.125	2.76
	AQRS	27	71	1.00	0.125	20.39
	WAQRS	23	29	1.00	0.125	52.65
	CQRS	24	55	1.00	0.125	512.1
$0.8N(0,1)+0.2N(10,1)$	SIS	49	79	1.00	0.125	1.04
	$\text{QaSIS}_{0.5}$	97	230	0.875	0.125	3.00
	$\text{QaSIS}_{0.75}$	217	343	0.75	0.125	2.78
	AQRS	37	75	1.00	0.125	21.83
	WAQRS	34	99	1.00	0.125	52.12
	CQRS	33	74	1.00	0.125	480.99

Table 3.3: Example 3: Dependent Model

$\boldsymbol{\varepsilon}$	Method	$\mathcal{R}_{0.5}$	$\mathcal{R}_{IQR}$	$\mathcal{S}_{0.5}$	$\mathcal{S}_{IQR}$	$T_{0.5}$
$N(0,1)$	SIS	26	53	1.00	0.125	1.065
	$\text{QaSIS}_{0.5}$	68	169	0.875	0.25	1.957
	$\text{QaSIS}_{0.75}$	99	193	0.875	0.125	1.96
	AQRS	31	69	1.00	0.125	15.413
	WAQRS	26	50	1.00	0.125	51
	CQRS	51	79	0.875	0.125	232.523
	WCQRS	52	136	0.875	0.125	410.451
t_1	SIS	873	166	0.125	0.00	0.69
	$\text{QaSIS}_{0.5}$	154	330	0.875	0.125	2.048
	$\text{QaSIS}_{0.75}$	245	362	0.75	0.25	2.051
	AQRS	130	235	0.875	0.125	15.492
	WAQRS	83	170	0.875	0.25	53.857
	CQRS	64	159	0.875	0.125	369.22
$0.9N(0,1)+0.1t_1$	SIS	40	72	0.875	0.125	0.669
	$\text{QaSIS}_{0.5}$	82	228	0.875	0.25	1.987
	$\text{QaSIS}_{0.75}$	146	241	0.75	0.125	1.99
	AQR	37	89	0.9375	0.125	16.015
	WAQRS	34	92	1.00	0.125	74.29
	CQRS	43	162	0.875	0.125	266.98
	WCQRS	37	64	1	0.125	515.68
$0.8N(0,1)+0.2t_1$	SIS	101	135	0.875	0.125	0.952
	$\text{QaSIS}_{0.5}$	119	237	0.875	0.125	1.986
	$\text{QaSIS}_{0.75}$	356	431	0.75	0.125	1.986
	AQRS	40	113	0.875	0.125	19.06
	WAQRS	36	58	1.00	0.125	52.95
	CQRS	58	193	0.875	0.125	240.67
	WCQRS	44	85	0.875	0.125	563.358
$0.9N(0,1)+0.1N(10,1)$	SIS	71	136	0.875	0.125	0.66
	$\text{QaSIS}_{0.5}$	119	237	0.875	0.125	1.984
	$\text{QaSIS}_{0.75}$	356	431	0.75	0.125	1.986
	AQRS	77	177	0.875	0.25	15.97
	WAQRS	63	193	0.875	0.125	54.156
	CQRS	57	144	0.875	0.125	251.34
$0.8N(0,1)+0.2N(10,1)$	SIS	131	246	0.875	0.125	0.66
	$\text{QaSIS}_{0.5}$	172	424	0.75	0.125	1.967
	$\text{QaSIS}_{0.75}$	428	434	0.625	0.125	1.967
	AQRS	115	243	0.875	0.125	16.29
	WAQRS	100	162	0.875	0.125	59.405
	CQRS	105	207	0.875	0.125	275.195

Table 3.4: Example 4: Heteroscedastic Model

$\boldsymbol{\varepsilon}$	Method	$\mathcal{R}_{0.5}$	$\mathcal{R}_{IQR}$	$\mathcal{S}_{0.5}$	$\mathcal{S}_{IQR}$	$T_{0.5}$
$N(0,1)$	SIS	9	10	1.00	0.00	0.749
	$\text{QaSIS}_{0.5}$	5	0	1.00	0.00	2.098
	$\text{QaSIS}_{0.75}$	15	33	1.00	0.125	2.102
	AQRS	9	5	1.00	0.00	15.599
	WAQRS	8	4	1.00	0.00	54.33
	CQRS	9	7	1.00	0.00	339.09
	WCQRS	8	7	1.00	0.00	405.026
t_1	SIS	907	185	0.00	0.125	1.057
	$\text{QaSIS}_{0.5}$	5	1	1.00	0.00	2.622
	$\text{QaSIS}_{0.75}$	131	368	0.75	0.25	2.624
	AQRS	54	190	0.875	0.25	15.66
	WAQRS	22	57	1.00	0.125	72.731
	CQRS	23	98	1.00	0.125	272.48
$0.9N(0,1)+0.1t_1$	SIS	317	751	0.625	0.625	0.982
	$\text{QaSIS}_{0.5}$	5	0	1.00	0.00	2.14
	$\text{QaSIS}_{0.75}$	19	36	1.00	0.125	2.147
	AQRS	10	20	1.00	0.00	15.703
	WAQRS	9	13	1.00	0.00	70.634
	CQRS	9	6	1.00	0.00	246.512
	WCQRS	9	24	1.00	0.00	721.401
$0.8N(0,1)+0.2t_1$	SIS	510	677	0.5	0.625	0.665
	$\text{QaSIS}_{0.5}$	5	0	1.00	0.00	1.988
	$\text{QaSIS}_{0.75}$	44	124	0.875	0.125	1.992
	AQRS	10	11	1.00	0.00	16.022
	WAQRS	9	16	1.00	0.00	71.086
	CQRS	9	8	1.00	0.00	274.141
	WCQRS	9	8	1.00	0.00	703.912
$0.9N(0,1)+0.1N(10,1)$	SIS	202	331	0.625	0.25	1.221
	$\text{QaSIS}_{0.5}$	5	0	1.00	0.00	2.101
	$\text{QaSIS}_{0.75}$	28	54	1.00	0.125	2.084
	AQRS	15	26	1.00	0.00	15.936
	WAQRS	13	26	1.00	0.031	51.917
	CQRS	11	20	1.00	0.00	280.303
$0.8N(0,1)+0.2N(10,1)$	SIS	358	555	0.375	0.5	0.677
	$\text{QaSIS}_{0.5}$	5	1	1.00	0.00	2.002
	$\text{QaSIS}_{0.75}$	104	154	0.875	0.25	2.008
	AQRS	91	251	0.875	0.344	15.761
	WAQRS	18	42	1.00	0.125	58.915
	CQRS	23	49	1.00	0.125	298.9

Table 3.5: Example 5: Additive Model ($\rho = 0$)

ε	Method	$\mathcal{R}_{0.5}$	$\mathcal{R}_{IQR}$	$\mathcal{S}_{0.5}$	$\mathcal{S}_{IQR}$	$T_{0.5}$
$N(0,1)$	NIS	683	429	0.50	0.00	25.13
	$\text{QaSIS}_{0.5}$	705	378	0.50	0.00	6.14
	$\text{QaSIS}_{0.75}$	757	415	0.50	0.25	5.62
	AQRS	625	392	0.50	0.00	26.41
	WAQRS	613	341	0.50	0.00	73.92
	CQRS	713	341	0.50	0.00	904.87
	WCQRS	663	410	0.50	0.00	1256.97
t_1	NIS	738	345	0.25	0.25	39.31
	$\text{QaSIS}_{0.5}$	700	419	0.50	0.00	5.13
	$\text{QaSIS}_{0.75}$	750	293	0.25	0.25	4.90
	AQRS	734	321	0.50	0.00	21.43
	WAQRS	646	451	0.50	0.00	77.17
	CQRS	730	362	0.50	0.00	922.43
$0.9N(0,1) + 0.1t_1$	NIS	696	387	0.50	0.00	36.88
	$\text{QaSIS}_{0.5}$	697	393	0.50	0.00	3.75
	$\text{QaSIS}_{0.75}$	654	422	0.50	0.25	3.64
	AQRS	715	365	0.50	0.00	29.48
	WAQRS	624	387	0.50	0.00	106.88
	CQRS	718	405	0.50	0.00	882.47
	WCQRS	668	455	0.50	0.00	934.80
$0.8N(0,1) + 0.2t_1$	NIS	720	353	0.50	0.00	36.48
	$\text{QaSIS}_{0.5}$	693	391	0.50	0.00	3.48
	$\text{QaSIS}_{0.75}$	696	386	0.50	0.25	3.38
	AQRS	723	311	0.50	0.00	29.73
	WAQRS	635	378	0.50	0.00	120.27
	CQRS	725	364	0.50	0.00	910.55
	WCQRS	687	389	0.50	0.00	1238.39
$0.9N(0,1) + 0.1N(10,1)$	NIS	696	368	0.50	0.00	36.21
	$\text{QaSIS}_{0.5}$	692	285	0.50	0.00	3.56
	$\text{QaSIS}_{0.75}$	619	343	0.50	0.25	3.34
	AQRS	628	386	0.50	0.00	21.21
	WAQRS	692	315	0.50	0.00	29.73
	CQRS	720	380	0.50	0.00	916.57
$0.8N(0,1) + 0.2N(10,1)$	NIS	699	413	0.50	0.00	35.37
	$\text{QaSIS}_{0.5}$	728	297	0.50	0.00	3.74
	$\text{QaSIS}_{0.75}$	724	336	0.50	0.25	3.52
	AQRS	703	364	0.50	0.00	20.95
	WAQRS	665	421	0.50	0.00	122.06
	CQRS	733	460	0.50	0.00	900.46

Table 3.6: Example 5: Additive Model ($\rho = 0.6$)

ε	Method	$\mathcal{R}_{0.5}$	$\mathcal{R}_{IQR}$	$\mathcal{S}_{0.5}$	$\mathcal{S}_{IQR}$	$T_{0.5}$
$N(0,1)$	NIS	84	405	0.75	0.25	35.57
	$\text{QaSIS}_{0.5}$	45	106	1.00	0.25	5.12
	$\text{QaSIS}_{0.75}$	36	117	1.00	0.25	4.67
	AQRS	16	46	1.00	0.00	24.22
	WAQRS	14	42	1.00	0.00	32.39
	CQRS	51	70	1.00	0.25	1180.76
	WCQRS	42	191	0.75	0.25	1221.39
t_1	NIS	272	639	0.75	0.75	35.55
	$\text{QaSIS}_{0.5}$	85	155	0.75	0.75	5.39
	$\text{QaSIS}_{0.75}$	58	92	1.00	1.00	5.08
	AQRS	49	49	1.00	1.00	31.61
	WAQRS	36	127	1.00	1.00	34.51
	CQRS	93	179	0.75	0.75	1277.32
$0.9N(0,1) + 0.1t_1$	NIS	130	319	0.75	0.75	26.02
	$\text{QaSIS}_{0.5}$	50	183	1.00	1.00	3.80
	$\text{QaSIS}_{0.75}$	40	84	1.00	1.00	3.68
	AQRS	27	62	1.00	1.00	24.37
	WAQRS	19	33	1.00	1.00	31.23
	CQRS	55	104	1.00	1.00	1216.99
	WCQRS	48	112	1.00	1.00	1353.21
$0.8N(0,1) + 0.2t_1$	NIS	181	496	0.75	0.75	26.09
	$\text{QaSIS}_{0.5}$	53	134	1.00	1.00	4.88
	$\text{QaSIS}_{0.75}$	46	73	1.00	1.00	4.66
	AQRS	31	49	1.00	1.00	24.52
	WAQRS	26	36	1.00	1.00	32.45
	CQRS	59	106	1.00	1.00	1196.22
	WCQRS	52	146	1.00	1.00	1291.70
$0.9N(0,1) + 0.1N(10,1)$	NIS	159	535	0.75	0.75	35.30
	$\text{QaSIS}_{0.5}$	66	157	0.874	0.875	5.75
	$\text{QaSIS}_{0.75}$	28	59	1.00	1.00	5.26
	AQRS	18	45	1.00	1.00	21.36
	WAQRS	16	72	1.00	1.00	24.41
	CQRS	57	117	1.00	1.00	920.50
$0.8N(0,1) + 0.2N(10,1)$	NIS	188	683	0.75	0.75	35.98
	$\text{QaSIS}_{0.5}$	77	279	0.75	0.75	6.79
	$\text{QaSIS}_{0.75}$	45	118	1.00	1.00	6.16
	AQRS	20	27	1.00	1.00	24.22
	WAQRS	17	61	1.00	1.00	30.25
	CQRS	84	120	0.75	0.75	1160.92

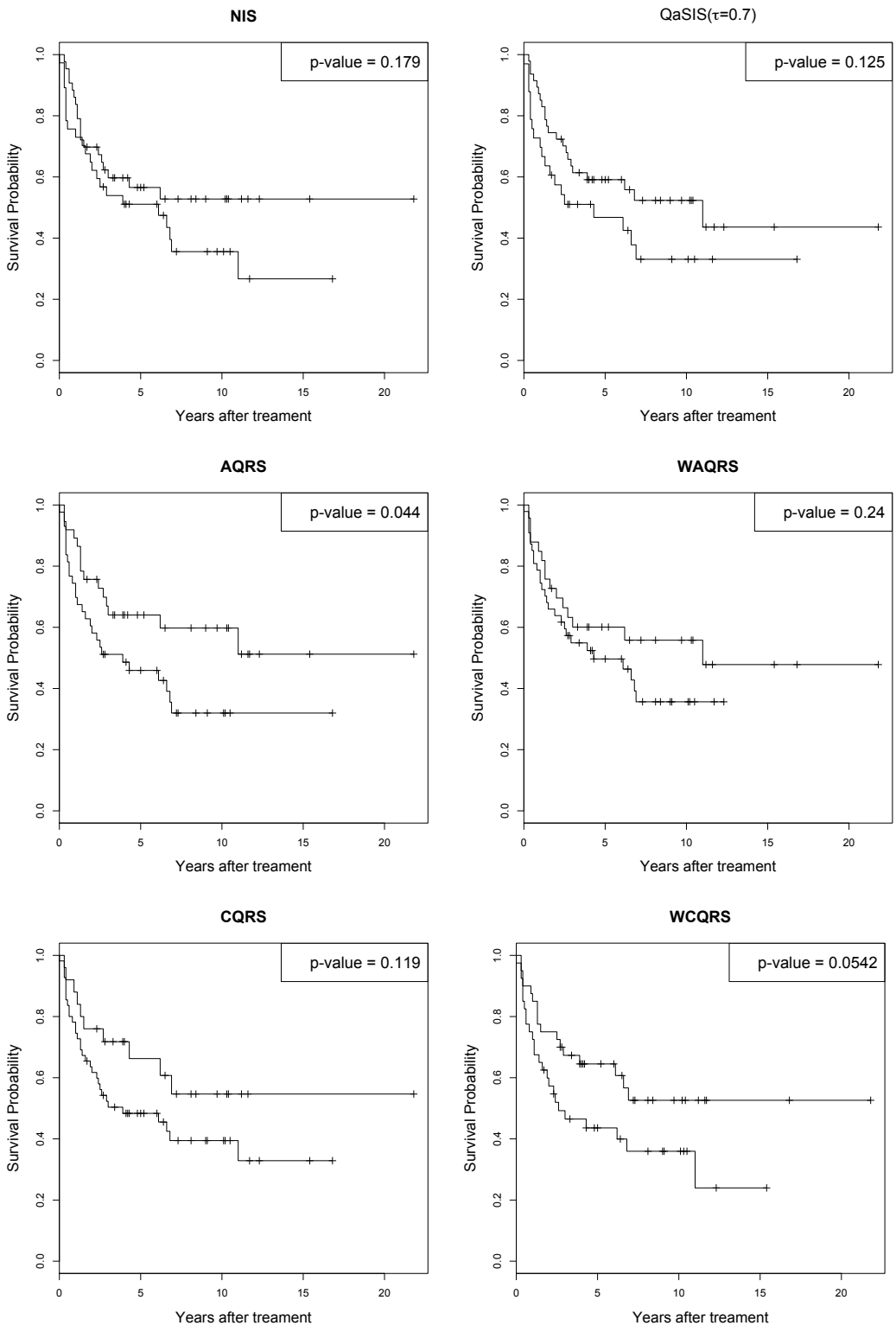

Figure 3.1: The Kaplan-Meier estimates of survival curves for the two risk groups in the testing data.

Chapter 4

Conclusion

4.1 Summary

Motivated by He, Wang and Hong [11], this thesis develops a more efficient variable screening procedure. We combine the information from multiple quantile levels by using the average quantile regression (AQR), weighted average quantile regression (WAQR), composite quantile regression (CQR), weighted composite quantile regression (WCQR) estimators to improve the efficiency. We develop the screening procedure in linear regression model and nonlinear regression model. In the nonlinear regression model, B-spline approximations are employed to estimate the conditional quantiles of Y. In addition, two different threshold rules are proposed and compared. We conduct simulation studies and a real data example to investigate the finite sample performance of the screening procedure based on the four estimators and compare them with SIS, NIS and QaSIS. We also study the difference between soft and hard threshold rules via a simulation.

From the results of the simulation and real data examples, we get the fol-

lowing conclusion: (i) SIS and NIS procedure exhibit the best performance when the random error has a normal distribution, but their performance deteriorate substantially for heavy-tailed or heteroscedastic errors. (ii) AQRS, WAQRS, CQRS and WCQRS significantly outperform QaSIS which considers only a single quantile level. (iii) For almost all the cases discussed, weighted methods outperform equally weighted ones. (iv) For non-normal distributions, our four approaches substantially perform better than SIS and NIS, whereas they are comparable for $N((0,1)$. (v) In practice, the weighted one is not necessarily better than the average one due to the difficulty in estimating $f(Q(\tau))$ properly.

4.2 Future Work

The following issues deserve further study:

- From He, Wang and Hong [11], we know QaSIS enjoys sure independence screening property, that is, all truly important predictors can be selected with probability approaching one as the sample size goes to infinity. Hence, we will show that our proposed procedure based on the four estimators also have that desirable property.

- When estimating $f(Q(\tau))$, we propose two different approaches. In the simulation and real data examples, we just choose the second one, which is simpler. From the results of the numerical studies, we see that it is really difficult to estimate it properly. Hence, the other method can be applied to estimate $f(Q(\tau))$.

- Taking the linear model as example, in the procedure based on the AQR

estimator, we define $D_{nj}^{\mathrm{AQR}} = K^{-1} \sum_{k=1}^{K} \left[Q_{\tau_k} \left(Y | X_j \right) - Q_{\tau_k} \left(Y \right) \right]$ and have the conclusion that it is expected to be close to zero if X_j is independent of Y. However, it is likely that there are some possible correlations between different quantiles. In this way, it is likely that the effects from different quantiles can be balanced out. Therefore, we may improve our AQRS method by defining D_{nj}^{AQR} in a different way

$$D_{nj}^{\mathrm{AQR}} = \frac{1}{K} \sum_{k=1}^{K} \left| Q_{\tau_k} \left(Y | X_j \right) - Q_{\tau_k} \left(Y \right) \right| .$$

Bibliography

[1] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology*, 2(4):511–522, 2004.

[2] J. Bradic, J. Fan, and W. Wang. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349, 2011.

[3] J. Fan. Comments on «Wavelets in statistics: A review» by A. Antoniadis. *Journal of the Italian Statistical Society*, 6(2):131–138, 1997.

[4] J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011.

[5] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[6] J. Fan and R. Li. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *International Congress of Mathematicians*, 3:595–622, 2006.

[7] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[8] J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038, 2009.

[9] J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

[10] P. Hall and H. Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3):533–550, 2009.

[11] X. He, L. Wang, and H. G. Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(1):342–369, 2013.

[12] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[13] D. R. Hunter and K. Lange. Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000.

[14] X. Jiang, J. Jiang, and X. Song. Oracle model selection for nonlinear models based on weighted composite quantile regression. *Statistica Sinica*, 22(4):1479–1506, 2012.

[15] B. Kai, R. Li, and H. Zou. Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):49–69, 2010.

[16] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, 28(5):1356–1378, 2000.

[17] R. Koenker. *Quantile regression*. Cambridge university press, 2005.

[18] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, 46(1):33–50, 1978.

[19] R. W. Koenker and V. d'Orey. Algorithm as 229: Computing regression quantiles. *Applied Statistics*, 36:383–393, 1987.

[20] G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877, 2012.

[21] H. Li and Y. Luan. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21(10):2403–2409, 2005.

[22] R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

[23] W. Lu and L. Li. Boosting method for nonlinear transformation models with censored survival data. *Biostatistics*, 9(4):658–667, 2008.

[24] X. Luo, L. A. Stefanski, and D. D. Boos. Tuning variable selection procedures by adding noise. *Technometrics*, 48(2):165–175, 2006.

[25] S. Portnoy, R. Koenker, et al. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300, 1997.

[26] A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3:267–288, 1996.

[28] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[29] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

[30] Y. Wu, D. D. Boos, and L. A. Stefanski. Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, 102(477):235–243, 2007.

[31] Y. Wu and Y. Liu. Variable selection in quantile regression. *Statistica Sinica*, 19(2):801–817, 2009.

[32] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[33] Z. Zhao and Z. Xiao. Efficient regressions via optimally combining quantile information. *Econometric Theory*, FirstView:1–43, 2014.

[34] L.-P. Zhu, L. Li, R. Li, and L.-X. Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475, 2011.

[35] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[36] H. Zou and M. Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3):1108–1126, 2008.